DOCUMENT RESUME

ED 419 023 TM 028 329

AUTHOR Thompson, Bruce
TITLE Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas.
PUB DATE 1998-04-00
NOTE 102p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Educational Research; *Effect Size; *Research Methodology; Scores; *Statistical Significance; Tables (Data); *Test Reliability
IDENTIFIERS Stepwise Regression; *Weighting (Statistical)

ABSTRACT

After presenting a general linear model as a framework for discussion, this paper reviews five methodology errors that occur in educational research: (1) the use of stepwise methods; (2) the failure to consider in result interpretation the context specificity of analytic weights (e.g., regression beta weights, factor pattern coefficients, discriminant function coefficients, canonical function coefficients) that are part of all parametric quantitative analyses; (3) the failure to interpret both weights and structure coefficients as part of result interpretation; (4) the failure to recognize that reliability is a characteristic of scores, and not of tests; and (5) the incorrect interpretation of statistical significance and the related failure to report and interpret the effect sizes present in all quantitative analysis. In several cases small heuristic discriminant analysis data sets are presented to make the discussion of each of these five methodology errors more concrete and accessible. Four appendixes contain computer programs for some of the analyses. (Contains 19 tables, 1 figure, and 143 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

aeraaddr.wp1 4/3/98

ED 419 023

Five Methodology Errors in Educational Research:

The Pantheon of Statistical Significance and Other Faux Pas

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Bruce Thompson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

I m03 8329

Invited address (Divisions E, D, and C) presented at the annual meeting (session #25.66) of the American Educational Research Association, San Diego, April 15, 1998. The assistance of Xitao Fan, Utah State University, in running the LISREL structural equation modeling program as the general linear model, is appreciated. The author may be contacted through Internet URL: **http://acs.tamu.edu/~bbt6147/.**

BEST COPY AVAILABLE

ABSTRACT

After presenting a general linear model as a framework for discussion, the present paper reviews five methodology errors that occur in educational research: (a) the use of *stepwise* methods; (b) the failure to consider in result interpretation the *context specificity* of analytic weights (e.g., regression beta weights, factor pattern coefficients, discriminant function coefficients, canonical function coefficients) that are part of all parametric quantitative analyses; (c) the failure to interpret *both weights and structure coefficients* as part of result interpretation; (d) the failure to recognize that *reliability* is a characteristic of scores, and not of tests; and (e) the incorrect interpretation of *statistical significance* and the related failure to report and interpret the *effect sizes* present in all quantitative analyses. In several cases small heuristic discriminant analysis data sets are presented to make more concrete and accessible the discussion of each of these five methodology errors.

A well-known popular cliche holds that a chain is only as strong as its weakest link. So, too, a research study will be at least partially compromised by whatever is the weakest link in the sequence of activities that cumulate in a completed investigation. Too often the weakest link in contemporary quantitative educational research involves the methodologies of statistical analysis.

There is no question that educational research, whatever its methodological and other limits, has influenced and informed educational practice (cf. Gage, 1985; Travers, 1983). But there seems to be some consensus that "too much of what we see in print is seriously flawed" as regards research methods, and that "much of the work in print ought not to be there" (Tuckman, 1990, p. 22). Gall, Borg and Gall (1996) concurred, noting that "the quality of published studies in education and related disciplines is, unfortunately, not high" (p. 151).

Empirical studies of published research involving methodology experts as judges corroborate these holistic impressions. For example, Hall, Ward and Comer (1988) and Ward, Hall and Schramm (1975) found that over 40% and over 60%, respectively, of published research was judged by methods experts as being seriously or completely flawed. Wandt (1967) and Vockell and Asher (1974) reported similar results from their empirical studies of the quality of published research. Dissertations, too, have been examined, and too often have been found methodologically wanting (cf. Thompson, 1988a, 1994a).

Of course, it must be acknowledged that even a methodologically flawed study may still contribute something to our

understanding of educational phenomena. As Glass (1979) noted, "Our research literature in education is not of the highest quality, but I suspect that it is good enough on most topics" (p. 12).

But the problem with methodologically flawed studies is that these methodological flaws are entirely *gratuitous*. There is no upside to conducting incorrect statistical analyses. Usually a more thoughtful analysis is not appreciably more demanding in time or expertise than is a compromised choice. Rather, incorrect analyses arise from doctoral methodology *instruction* that teaches research methods as series of rotely-followed routines, as against thoughtful elements of a reflective enterprise; from doctoral *curricula* that seemingly have less and less room for quantitative statistics and measurement content, even while our knowledge base in these areas is burgeoning (Aiken, West, Sechrest, Reno, with Roediger, Scarr, Kazdin & Sherman, 1990; Pedhazur & Schmelkin, 1991, pp. 2-3); and, in some cases, from an unfortunate *atavistic impulse* to somehow escape responsibility for analytic decisions by justifying choices, sans rationale, solely on the basis that the choices are common or traditional.

Purpose of the Paper

The purpose of the present paper is to review five methodology errors that occur in educational research: (a) the use of *stepwise* methods; (b) the failure to consider in result interpretation the *context specificity* of analytic weights (e.g., regression beta weights, factor pattern coefficients, discriminant function coefficients, canonical function coefficients) that are part of all

parametric quantitative analyses; (c) the failure to interpret *both weights and structure coefficients* as part of result interpretation; (d) the failure to recognize that *reliability* is a characteristic of scores, and not of tests; and (e) the incorrect interpretation of *statistical significance* and the related failure to report and interpret the *effect sizes* present in all quantitative analyses. These comments are not new to the literature, or even to my own writing. But the field has seemingly remained somewhat recalcitrant in reflecting evolution as regards these methodological issues.

The paper presents a conceptual overview of each concern. In several cases small heuristic data sets are presented to make more concrete and accessible the discussion of each of these five methodology errors. Because, as will be shown, all parametric methods are part of one general linear model (GLM) family, methodology dynamics illustrated for one heuristic example generalize to other related cases. In the present paper, discriminant analysis examples are consistently (but arbitrarily) employed as heuristics. Nevertheless, the illustrations necessarily generalize to other analyses within the GLM family.

Delimitation

Of course, methodological errors other than these five might have been cited. For example, empirical studies (Emmons, Stallings & Layne, 1990) show that, "In the last 20 years, the use of multivariate statistics has become commonplace" (Grimm & Yarnold, 1995, p. vii), probably for very good reasons (Fish, 1988; Thompson, 1984, 1994e). Many such studies employ MANOVA (all to the

good), but an unfortunate number of these studies then use ANOVA methods post hoc to explore detected multivariate effects (all to the bad) (Borgen & Seling, 1978). As I have noted elsewhere,

> The multivariate analysis evaluates multivariate synthetic variables, while the univariate analysis only considers univariate latent variables. Thus, *univariate post hoc tests do not inform the researcher about the differences in the multivariate latent variables actually analyzed in the multivariate analysis...* It is illogical to first declare interest in a multivariate omnibus system of variables, and to then explore detected effects in this multivariate world by conducting non-multivariate tests! (Thompson, 1994e, p. 14, emphasis in original)

Similarly, all too often researchers erroneously interpret the eigenvalues in factor analysis as reflecting the variance contained in the individual factors *after* rotation (Thompson & Daniel, 1996a). Or the discarding of variance in order to conduct ANOVA (cf. Thompson, 1985) or incorrect use of ANCOVA (Thompson, 1992b) might have been discussed. However, space precludes discussion here of all possible common methodology errors; the present discussion necessarily must be delimited in some manner.

Premise Regarding Movement in Fields

In considering these five methodology errors, it may be important for each of us to remember that, over the course of careers, fields, including the methodology-related fields, do move.

Invariably, those of us in the late stages of our careers will confront the realization that some methodology choices in our own work, published decades earlier, no longer reflect standards of present best practice, or might even now be deemed fully inappropriate. Responsible scholars must remain open, and be willing to engage in continual reflection as to whether our own personal analytic traditions remain viable.

Some have suggested that resistance to adopting revised methodological practice may in some cases be an artifact of denial, cognitive dissonance, and other classical psychological dynamics (Thompson, in press-d). For example, Schmidt and Hunter (1997) noted that "changing the beliefs and practices of a lifetime... naturally... provokes resistance" (Schmidt & Hunter, 1997, p. 49). Similarly, Rozeboom (1960) observed that "the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

Recognizing the reality that fields move, and that to be fair works must be evaluated primarily against the methodological standards contemporary at the time of a given report, may facilitate helpful change. Prior to advocating selected changes, however, the general linear model (GLM) will be briefly described so as to provide a unifying conceptual framework for the remaining discussion. Structural equation modeling (SEM) will be presented as the most general case of the general linear model (GLM).

Conceptual Framework: SEM as the General Linear Model (GLM)

In one of his innumerable seminal contributions, the late

Jacob "Jack" Cohen (1968) demonstrated that multiple regression subsumes all the univariate parametric methods as special cases, and thus provides a univariate general linear model that can be employed in all univariate analyses. Ten years later, in an equally important article Knapp (1978) presented the mathematical theory showing that canonical correlation analysis subsumes all the parametric analyses, both univariate and multivariate, as special cases. More concrete demonstrations of these relationships have also been offered (Fan, 1996; Thompson, 1984, 1991, in press-a). Both the Cohen (1968) and the Knapp (1978) articles were cited within a compilation of the most noteworthy methodology articles published during the last 50 years (Thompson & Daniel, 1996b).

However, structural equation modeling (SEM) represents an even bigger conceptual tent subsuming more restrictive methods (Bagozzi, 1981). Instructive illustrations of these relationships have been offered by Fan (1997). Prior to extracting the conceptual implications of the realization that a general linear model underlies all parametric analyses, a concrete demonstration that SEM is a general linear model subsuming canonical correlation analysis (CCA) (and its multivariate and univariate special cases) may be useful.

Heuristic Illustration that SEM Subsumes CCA

The illustration that SEM is a general linear model subsuming canonical correlation analysis (and its multivariate and univariate special cases) employs scores on seven variables (i.e., two in one set, and three in the other set) from the 301 cases in the Holzinger and Swineford (1939, pp. 81-91) data. These scores on

ability batteries have classically been used as examples in both popular textbooks (Gorsuch, 1983, *passim*) and computer program manuals (Jöreskog & Sörbom, 1989, pp. 97-104), and thus are familiar to many readers.

Table 1 presents the bivariate correlation matrix for these data. As in all parametric analyses, a correlation or covariance matrix is the basis for all analyses; this matrix is partitioned into quadrants (see Table 1) honoring the variables' membership in criterion or predictor sets, and is then subjected to a principal components analysis (Thompson, 1984, in press-a).

INSERT TABLE 1 ABOUT HERE.

Appendix A presents the SPSS/LISREL computer program used to analyze the data. Table 2 presents the SPSS canonical correlation analysis of these same data.

INSERT TABLE 2 ABOUT HERE.

Table 3 presents the relevant portions of the LISREL analysis of the canonical correlation model for these data. The LISREL coefficients for the "gamma" matrix exactly match (within rounding error) the SPSS canonical function coefficients presented in Table 2. The only exception is that all the signs for the SEM second canonical function coefficients must be "reflected." "Reflecting" a function (changing *all* the signs on a given function, factor, or equation) is *always* permissible, because the scaling of psychological constructs is arbitrary. Thus, the SEM and the canonical analysis derived the same results. Since SEM can be

employed to test a CCA model, SEM is an even more general case of the general linear model, *quod erat demonstrandum*.

INSERT TABLE 3 ABOUT HERE.

Heuristic Implications

There are a number of implications that can be drawn from the realization that a general linear model subsumes other methods as special cases. Specifically, all classical parametric methods are least squares procedures that implicitly or explicitly (a) use least squares weights (e.g., regression beta weights, standardized canonical function coefficients) to optimize explained variance and minimize model error variance, (b) focus on latent synthetic variables (e.g., the regression $\hat{Y}$ variable) created by applying the weights (e.g., beta weights) to scores on measured/observed variables (e.g., regression predictor variables), and (c) yield variance-accounted-for effect sizes analogous to r^2 (e.g., R^2, eta^2, $omega^2$). Thus, all classical analytic methods are correlational (Knapp, 1978; Thompson, 1988a).

Designs may be experimental or correlational, but *all analyses are correlational*. Thus, an effect size analogous to r^2 can be computed in any parametric analysis (see Snyder and Lawson (1993), or Kirk (1996)).

The fact that all classical parametric methods use weights to then compute synthetic/latent variables by applying the weights to the measured/observed variables is obscured by the fact that most *computer packages* do not print the least squares weights that are actually invoked in ANOVA, for example, or when t-tests are

conducted. Thus, some researchers unconsciously presume that such methods do not invoke optimal weighting systems.

The fact that all classical parametric methods use weights to then compute synthetic/latent variables by applying the weights to the measured/observed variables is also obscured by the inherently confusing *language* of statistics. As I have noted elsewhere, the weights in different analyses

> ...are all analogous, but are given different names in different analyses (e.g., beta weights in regression, pattern coefficients in factor analysis, discriminant function coefficients in discriminant analysis, and canonical function coefficients in canonical correlation analysis), mainly to obfuscate the commonalities of [all] parametric methods, and to confuse graduate students. (Thompson, 1992a, pp. 906-907)

If all standardized weights across analytic methods were called by the same name (e.g., beta weights), then researchers might (correctly) conclude that all analyses are part of the same general linear model.

Indeed, both the weight systems (e.g., regression equation, factor) and the synthetic variables (e.g., the regression $\hat{Y}$ variable) are also arbitrarily given different names across the analyses, again mainly so as to confuse the graduate students. Table 4 summarizes some of the elements of the very effective conspiracy.

INSERT TABLE 4 ABOUT HERE.

The present paper will employ this general linear model as a unifying conceptual framework for some of the arguments made herein. However, prior to presenting these views, a brief digression is required.

Predictive Discriminant Analysis (PDA) as a Hybrid GLM Offshoot

In the seminal work on discriminant analysis, Huberty (1994; see also Huberty and Barton (1989) and Huberty and Wisenbaker (1992)) thoughtfully distinguished two major applications: descriptive discriminant analysis (DDA) and predictive discriminant analysis (PDA). Put simply, DDA *describes* the differences on intervally-scaled "response" variables associated with a nominally-scaled variable, membership in different groups. PDA, on the other hand, uses intervally-scaled "response" variables to *predict* membership in different groups. Thus, the purpose of the analysis distinguishes the two methods (and these purposes subsequently determine which aspects of the results are relevant or irrelevant).

The drawing of a distinction between DDA and PDA is not mere statistical nit-picking. Instead, the relevant aspects of DDA and PDA results are completely different. For example, in PDA the "hit rate" (and which response variables most contribute to the hit rate) is the *sina qua non* of the analysis, while the weights are generally irrelevant as regards result interpretation. In DDA, on the other hand, the weights and the "structure" of the synthetic/latent variable scores are very important to interpretation, but the concept of hit rate becomes irrelevant.

The number of systems of weights (i.e., "functions," or "rules") also differs across DDA and PDA. In DDA, the number of linear discriminant functions (LDFs) is the number of groups minus one, or the number of response variables, whichever is smaller. In PDA, the number of linear classification functions (LCFs) is the number of groups. For example, with two groups and three response variables, in DDA there would be one LDF (and an associated set of scores on the synthetic variable, the discriminant scores). In the same case, in PDA there would be two LDFs (and associated sets of scores on the synthetic variables, the classification scores).

PDA is a hybrid offshoot of the general linear model, while DDA resides fully within the GLM nuclear family. Thus, the conclusions reached here based on GLM concepts may not apply to the PDA case.

When More Variables Can Hurt Study Effects

One powerful demonstration of PDA versus DDA dynamics involves a paradox. In any GLM analysis, more variables (e.g., more regression predictors) always lead to effect sizes (e.g., R^2) that are equal to or greater than the effects associated with fewer variables. However, in PDA, more response variables can actually *hurt* the PDA hit rate.

The Table 5 data, drawn from the Holzinger and Swineford (1939) data described previously, can be analyzed to illustrate these dynamics. The Appendix B SPSS program conducts the relevant analyses.

INSERT TABLE 5 ABOUT HERE.

Table 6 presents the hit rates derived using *three* response variables as predictors using both LDF and LCF scores; these hit rates are both 66.4% ([40 + 31] / 107). [Normally only LCFs are used for classification purposes, even though SPSS incorrectly uses LDF scores for this purposes (Huberty & Lowman, 1997)]. Table 6 also presents the hit rates derived using *four* response variables as predictors using both LDF and LCF scores; these hit rates are both 63.6% ([38 + 30] / 107). Figure 1 presents the corresponding results in graphic form.

INSERT TABLE 6 AND FIGURE 1 ABOUT HERE.

Indeed, the hit rate differences with the use of three versus four response variables is even greater than the apparent difference of 71 versus 68 people, respectively, being correctly classified. In fact, as noted in Table 7, 9 persons were classified differently across the analyses using three versus four response variables, even though the net impact of using more predictors was a net loss in predictive accuracy of three hits. [If the same data were treated as reflecting a DDA case, the Wilks lambda effect size would be the same or better (i.e., a smaller lambda value) for four (0.8050684) as against three (0.8094909) response variables, as is always true in the GLM case.]

INSERT TABLE 7 ABOUT HERE.

Elsewhere I (Thompson, 1995b) have explained some of these counterintuitive dynamics by portraying a hypothetical set of results involving five response variables. Presume there were three

"fence-riders," that is, cases very near the classification boundaries (arbitrarily cases #4, #11, and #51). Let's say with five predictor variables our initial lambda is .50, and let's say we add an additional, sixth response variable as a PDA predictor.

Clearly, having more predictive information always help us better explain data dynamics, or at least can't take away what we already know. This is reflected by the fact that the Wilks lambda value will always stay the same or get better (i.e., smaller) as we add predictor variables.

> But this occurs only *on the average*, as reflected in on-the-average statistics such as lambda. While relative explanatory power will remain the same or improve on the average, at the case level each and every single case will not necessarily move toward its actual group's location when the additional sixth predictor variable is used. For example, let's say that all cases' positions except cases #4, #11, #51 and #43 remain fixed in essentially their initial locations and that group territorial boundaries also remain roughly unchanged.
>
> If because the sixth predictor was especially useful in locating case #43, case #43 might move very far toward but not over the boundary that would have yielded a correct classification. Lambda would reflect this change by getting better (i.e., smaller), such as changing from .50 to perhaps .45.

> Cases #4, #11, and #51 might move slightly away from their actual group, because although the sixth predictor will either not change explanatory power or will provide more information on the average, it is still possible that the sixth predictor may provide misinformation about these three particular cases, resulting in their moving across their actual group boundary and becoming misclassified. This small movement will, of course, be reflected in lambda, which will correspondingly get only slightly worse (i.e., bigger), such as moving from .45 to .46. Yet even though on the average locations have gotten more accurate and lambda has consequently improved from the original .50 to the final .46, the number of cases correctly classified when using all six predictors will have gotten worse by a net classification-accuracy change of minus three cases. (Thompson, 1995b, p. 345, emphasis in original)

Error #1: Using Stepwise Methods

Huberty (1994) has noted that, "It is quite common to find the use of 'stepwise analyses' reported in empirically based journal articles" (p. 261). Huebner (1991, 1992) and Jorgenson, Jorgenson, Gillis and McCall (1993) are a few examples from among the many egregious reports of stepwise analyses.

Stepwise methods continue to be used, notwithstanding scathing indictments of many of these applications (cf. Huberty, 1989; Snyder, 1991). My own feelings are intimated by the title of one of

my editorials, viz. "Why won't stepwise methods die?" (Thompson, 1989).

Three major problems with stepwise can be noted, and will be briefly summarized here. A more complete treatment is available in Thompson (1995c).

The consequences of these three problems are quite serious. As Cliff (1987, p. 185) noted, "most computer programs for [stepwise] multiple regression are positively satanic in their temptations toward Type I errors." He also suggested that, "a large proportion of the published results using this method probably present conclusions that are not supported by the data" (pp. 120-121).

Wrong Degrees of Freedom

First, *most computer packages (and thus most researchers) use the wrong degrees of freedom in their statistical significance tests for stepwise methods, thus systematically always inflating the likelihood of obtaining statistically significant results.* Degrees of freedom are the "coins" we pay to investigate the dynamics within our data. The statistical significance tests take into account both the number of coins we've chosen to spend and the number we have chosen to reserve.

The most rigorous tests occur when we spend few degrees of freedom and reserve many. Conversely, at the extreme, all models with no degrees of freedom reserved (i.e., degrees of freedom error =0) always fit the data perfectly. For example, the bivariate r^2 with n=2 inherently is always 1.0, as long as both X and Y are variables. Similarly, the multiple regression R^2 with two predictors variables and n=3 inherently must always be 1.0.

The computer packages conventionally charge degrees of freedom for the numerator (synonymously also called "model," "between," "regression," and "explained," to confuse the graduate students) that are a function of the number of response variables "entered" in the analysis at a given step. The remaining degrees of freedom (synonymously called "denominator," "residual," "error," "within," and "unexplained") are inversely related to the number of response variables "entered" in a given step.

Table 8 illustrates these dynamics for a study involving 2 steps of stepwise analysis, with k=3 groups and n=120 people. Table 8 compares the results for two steps of analysis using the degrees of freedom calculations employed by SPSS and other computer packages, labelled "Incorrect," with the same calculations employing the correct degrees of freedom.

INSERT TABLE 8 ABOUT HERE.

The differences in the analyses revolves around what "entered" means. The computer packages define "entered" or "used" as actually entered into the prediction equation. Thus, in step one the packages consider that only one predictor has been entered, while in step two the packages consider that two response variables have been entered.

However, in this example each and every one of the 50 response variables was "used" at each and every one of the three steps, to decide which variable to enter at each step. The 49 or 48 unselected response variables may not have been retained in the analysis, but each one was examined, and played with, and actually

tasted, prior to the leftovers then being returned to the cafeteria display case.

This system of determining the degrees of freedom bill is analogous to only charging John Belushi in the movie Animal House for the food on his cafeteria tray, and charging nothing for what he has tasted and discarded. Clearly, this statistical package system of coinage is wrong. [Charging only for variables actually entered at each step would be appropriate, for example, if these response variables were randomly selected without first tasting each and every response variable.]

It is instructive to see how using the wrong degrees of freedom in the numerator of the statistical significance testing calculations, and the wrong denominator df in the calculations, both bias the tests in favor of getting statistical significance. Table 8 illustrates how dramatic the effect of using the wrong degrees of freedom can be.

After one step, the computer calculates that $F_{(2,117)} = 15.29841$, with an associated probability of .0000012; the correct $F_{(100,136)}$ is 0.16751, with an associated probability of 1.00000. After the second step, the computer calculates that $F_{(4,232)} = 13.64322$, with an associated probability of .0000945; the correct $F_{(100,136)}$ is 0.31991, with an associated probability of 1.00000. Obviously, the example illustrates that the correct and incorrect results can be night-vs-day different!

Three factors determine exactly how egregiously the use of the wrong degrees of freedom distorts the stepwise results. The distortions are increasingly serious as (a) sample size is smaller,

(b) the number of steps is larger, and (c) the number of response variables available to be selected is larger.

Nonreplicability of Results

Second, *stepwise methods tend to yield results that are sample-specific and do not generalize well to future studies.* This is because stepwise requires a linear sequence of decisions, each of which is contingent upon all the previous decisions in the sequence. This is very much like walking through a maze--an incorrect decision at any point will lead to a cascade of subsequent decisions that each may themselves be wrong.

Stepwise considers all differences of any magnitudes between variance explained by the response variables to be exact and true. Since there are usually numerous combinations of the response variables, and credit for variance explained for each partition of the variables may be influenced by sampling error, any small amount of sampling error anywhere in a single response variable can lead to disastrously erroneous choices in the linear sequence of stepwise selection decisions.

Stepwise Does NOT Identify the Best Variable Set of a Given Size

Third, *stepwise methods do not correctly identify the best set of predictors of a given response variable set size, k.* For example, if one has 30 response variables, and does three steps of analysis, it is possible that the best predictor set of size $k=3$ will include none of the three variables selected after three steps of stepwise analysis of the same data, and that the three stepwise variables would also yield a lower effect size.

This may seem counter-intuitive, but upon reflection, it

should be easy to see that in fact stepwise analysis does not seek to identify the best variable set of a certain size. Stepwise simply does not ask the question, "What is the best predictor set of a given size?" This question requires simultaneously considering all the combinations of the variables that are possible for a given set size. Stepwise analysis never *simultaneously* considers all the combinations of the predictor variables. Rather, at each step stepwise analysis takes the previously entered variables as a given, and then asks which *one* change in the predictor set will most improve the prediction.

Picking the best new variable in a sequence of selections is not the same as picking the best variable set of a given size. As Thompson (1995c) explained:

> Suppose one was picking a basketball team consisting of five players. The *stepwise* selection strategy picks the best potential player first, then the second best player in the context of the characteristics of the previously-selected first player, and so forth.
>
> An alternative strategy is an *all-possible-subsets* approach which asks, "which five potential players play together best as a team?". This team might conceivably contain exactly zero of the five players selected through the stepwise approach. Furthermore, this "best team" might be able to stomp the "stepwise team" by a considerable margin, because teams consisting of players of lesser

> abilities may still play together better as a team than players selected through a linear sequence of stepwise decisions. (pp. 528, 530, emphasis in original)

The Table 9 data provide a powerful heuristic. Table 10 presents an abridged printout for these data involving two steps of stepwise DDA, conducted using the Appendix C SPSS program. In this analysis the stepwise algorithm selects response variables X1 and X2, and the lambda value is .6553991 ($F_{(4,232)}$=13.64322).

INSERT TABLES 9 AND 10 ABOUT HERE.

Compare the Table 10 results with those in Table 11. Table 11 presents the DDA results for all six possible combinations of the four response variables considered two at a time. Note that the best set of two variables (i.e., smallest lambda) involves response variables X3 and X4 (λ = .6272538, $F_{(4,232)}$=15.23292). The best variable set of size two contained neither of the two variables selected by the stepwise analysis!!!!!

INSERT TABLE 11 ABOUT HERE.

Error #2: Ignoring the *Context Specificity* of GLM Weights

As noted previously, all univariate and multivariate methods apply weights to the measured variables to derive scores on the latent or synthetic variables that are actually the focus of all analyses. Consequently, if (and only if) noteworthy effects (e.g., R^2, Rc^2) are detected, it then becomes reasonable to consult the

weights as part of the process of determining which response variables contributed to the detected effect. Indeed, some researchers have even taken the view that these weights (e.g., beta weights, standardized discriminant function coefficients) should be the sole basis for evaluating the importance of response variables (Harris, 1989).

Unfortunately, overinterpretation of GLM weights is a serious threat. The weights can be greatly influenced by which variables are included or are excluded from a given analysis. Furthermore, Cliff (1987, pp. 177-178) noted that weights for a given set of variables may vary widely across samples, and yet consistently still yield the same effect sizes (i.e., be what he called statistically "sensitive"). Clearly weights are not the sole story in interpretation.

Any interpretations of weights must be considered *context-specific*. Any change in the variables in the model can radically alter <u>all</u> of the weights. Too few researchers appreciate the potential magnitudes of these impacts.

The Table 12 data illustrate these dynamics. The analysis contrasts using DDA models with either three response variables (i.e., X1, X2, and X3) or four response variables (i.e., X1, X2, X3, and X4). The example can be framed as either adding one response variable to an analysis involving three response variables, or deleting one response variable from an analysis involving four. This DDA example involves variance-covariance matrices for each of three groups that are exactly equal (called "homogeneity"), so the results are not confounded by failure to

meet one of the assumptions of the analysis.

INSERT TABLE 12 ABOUT HERE.

Table 13 presents an excerpt from an SPSS analysis of the Table 12 data conducted using the Appendix D computer program. Note the dramatic changes in the DDA standardized function coefficients. For example, with three response variables the first response variable, X1, had standardized function coefficients of 1.50086 and -.01817 on the two DDA functions. With four response variables X1 had standardized function coefficients of -.47343 and 1.22249 on the two DDA functions. Thus, the coefficients were quite variable in both magnitude and sign.

INSERT TABLE 13 ABOUT HERE.

These fluctuations are not problematic, if (and only if) the researcher has selected *exactly* the right model (i.e., has not made what statisticians call a model specification error). But as Pedhazur (1982) has noted, "The rub, however, is that the true model is seldom, if ever, known" (p. 229). And as Duncan (1975) has noted, "Indeed it would require no elaborate sophistry to show that we will never have the 'right' model in any absolute sense" (p. 101).

In other words, as a practical matter, the context-specificity of weights is always problematic, and the weights consequently must be interpreted cautiously. Some researchers acknowledge the vulnerability of the weights to sampling error influences (i.e., the so-called "bouncing beta" problem), but a more obvious concern

is the context-specificity of the weights in the real-world context of full or partial model misspecification.

Error #3: Failing to Interpret Both Weights and Structure Coefficients

A response variable given a standardized weight of zero is being obliterated by the multiplicative weighting process, indicating either that (a) the variable has zero capacity to explain relationships among the variables or that (b) the variable has some explanatory capacity, but one or more other variables yield the same explanatory information and are arbitrarily (not wrongly, just arbitrarily) receiving all the credit for the variable's predictive power. Because a response variable may be assigned a standardized multiplicative weight of zero when (b) the variable has some explanatory capacity, but one or more other variables yield the same explanatory information and are arbitrarily (not wrongly, just arbitrarily) given all the credit for the variable's predictive power, it is *essential* to evaluate other coefficients in addition to standardized weights during interpretation, to determine the specific basis for the weighting.

Just as it would be incorrect to evaluate predictor variables in a regression analysis only by consulting beta weights (Cooley & Lohnes, 1971, p. 55; Thompson & Borrello, 1985), in any GLM analysis it would be inappropriate to only consult standardized weights during result interpretation (Borgen & Seling, 1978, p. 692; Kerlinger & Pedhazur, 1973, p. 344; Levine, 1977, p. 20; Meredith, 1964, p. 55, Thompson, 1997b). Yet, some researchers do exactly that (cf. Humphries-Wadsworth, 1998).

Under most circumstances standardized weights are not correlation coefficients. Thus, some of the weights in the Table 11 are less than -1 or are greater than +1. Structure coefficients, on the other hand, are always correlation coefficients, and reflect the linear relationship between scores on a given measured or observed variable with the scores on a given latent or synthetic variable. Thus, because synthetic variable are actually the focus of all parametric analyses, and because structure coefficients reveal the structure of these latent variables, the importance of structure coefficients seems obvious.

Three possible cases can be delineated. The three illustrations demonstrate that jointly considering both standardized weights and structure coefficients indicates to the researcher which case is present in a given analysis. Appendix E presents the SPSS computer program used to analyze the three heuristic data sets.

Case #1: Function and Structure Coefficients are Equal

In the special GLM case where measured variables are uncorrelated, the standardized weights in this case (and in this case only) are correlation coefficients. For example, in regression, if the predictor variables are uncorrelated, each predictor variable's beta weight equals that variable's product-moment correlation with the criterion variable. In discriminant analysis, the same principle applies if the "pooled" correlation matrix of the response variables indicates that the response variables are uncorrelated.

Table 14 presents a hypothetical DDA data set illustrating this case for a k=3 group problem involving scores of n=30 people on each of p=3 response variables. As indicated by the Table 15 excerpt from the SPSS output for these data, in this special case the standardized function coefficients exactly equal the respective structure coefficients of the response variables.

INSERT TABLES 14 AND 15 ABOUT HERE.

Case #2: Measured Variables with Near-zero Weights Still Important

As noted previously, measured variables may be assigned multiplicative weights of zero if the measured variable contains useful variance, but that variance is also present in some combination of the other measured variables. The researcher interpreting these results, especially if only standardized weights are interpreted, might erroneously conclude that such a response variable with a near-zero weight had essentially no utility in generating the observed effect. Instead, the result merely indicates that this variable is arbitrarily being denied credit for its potential contributions.

Table 16 presents a relevant heuristic DDA data set for this case involving k=3 groups and p=3 response variables. Table 17 presents an excerpt from the related SPSS analysis of the tabled data.

INSERT TABLES 16 AND 17 ABOUT HERE.

In this example, the standardized function coefficient on

Function I for X3 was -.05507, while on the same function the other two response variables had standardized function coefficients of roughly +.95. Yet the squared structure coefficient (r_s^2 = .81431^2 = 66.3%) for X3 on the function indicates that X3 had more than twice the explanatory power as variables X1 (r_s^2 = .54141^2 = 29.3%) and X2 (r_s^2 = .56453^2 = 31.9%). Clearly, consulting only the function coefficients for this example would have resulted in a serious misinterpretation of results.

Case #3: "Suppressor" Effects

The previous case makes clear that a measured variable assigned a zero or near-zero weight may nevertheless be an important variable, as reflected in the variable having a large non-zero structure coefficient. However, although it may seem counter-intuitive, a measured/observed variable may also have a zero or near-zero structure coefficient, and still be very important in defining a detected effect, as reflected in the variable having a non-zero standardized weight. [That is, only measured variables with both near-zero weights and near-zero structure coefficients are useless in defining a given detected effect.]

Such a variable is classically termed a "suppressor" variable. However, although the name may feel pejorative, a "suppressor" variable actually increases the effect size, and so suppression is a good (and not a bad) thing. As defined by Pedhazur (1982, p. 104), in the related regression case, "A suppressor variable is a variable that has a zero, or close to zero, correlation with the

criterion but is correlated with one or more than one of the predictor variables." Henard (1998) provides a nice overview of suppressor effects.

Suppressor effects are quite difficult to explain in an intuitive manner. But Horst (1966) gave an example that is relatively accessible. He described the multiple regression prediction of pilot training success during World War II using mechanical, numerical, and spatial ability scores, each measured with paper and pencil tests. The verbal scores had very low correlations with the dependent variable, but had larger correlations with the other two predictors, since they were all measured with paper and pencil tests, i.e., measurement artifacts inflate correlations among traits measured with similar methods. As Horst (1966, p. 355) noted, "Some verbal ability was necessary in order to understand the instructions and the items used to measure the other three abilities."

Including verbal ability scores in the regression equation in this example actually served to remove the contaminating influence of one predictor from the other predictors, which effectively increased the R^2 value from what it would have been if only mechanical, numerical and spatial abilities had been used as predictors. The verbal ability variable had negative beta weights in the equation. As Horst (1966, p. 355) noted, "To include the verbal score with a negative weight served to suppress or subtract irrelevant ability, and to discount the scores [on the other predictors] of those who did well on the test simply because of

their verbal ability rather than because of abilities required for success in pilot training." The fact that a measured variable unrelated to a measured criterion variable can still make important contributions in an analysis itself makes the very important point that the latent or synthetic variables analyzed in all parametric methods are always more than the sum of their constituent parts.

Table 18 presents a relevant heuristic DDA data set for this case involving k=3 groups and p=3 response variables. Table 19 presents an excerpt from the related SPSS analysis of the tabled data. As reported in Table 19, on Function I DDA response variable X3 had a near-zero structure coefficient (r_s = -.03464), but a large non-zero standardized function coefficient (i.e., -1.58393). Indeed, on this function X3 had the largest absolute standardized function coefficient, since X1 and X2 had standardized function coefficients of +1.22956 and +1.21174, respectively.

INSERT TABLES 18 AND 19 ABOUT HERE.

Error #4: Failing to Recognize that Reliability Is *Not* a Characteristic of Tests

Nature of Score Reliability

Misconceptions regarding the nature of reliability abound within the social sciences. For example, some researchers do not realize that, "Notwithstanding erroneous folkwisdom to the contrary, sometimes scores from shorter tests are more reliable than scores from longer tests" (Thompson, 1990, p. 586). In her important recent article, Vacha-Haase (1998a) cited the example of

the Bem Sex-Role Inventory, noting that, "[i]n fact, the 20-item short-form of the Bem generally yields more reliable scores (r_{xx}^2 for the feminine scale ranging from .84 to .87) than does the 40-item long-form (r_{xx}^2 for the feminine scale ranging from .75 to .78)" (pp. 9-10).

Misconceptions regarding reliability flourish in part because

> [a]lthough most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures... (Pedhazur & Schmelkin, 1991, pp. 2-3)

Empirical study of doctoral curricula confirms this impression (Aiken et al., 1990).

The most fundamental problem is that too few researchers act on a conscious recognition that *reliability is a characteristic of scores or the data in hand*, and not of tests. Test booklets are not impregnated with reliability during the printing process. The WISC that yields reliable scores for some adults on a given occasion of measurement will not necessarily do so when the same test is administered to first-graders.

Many researchers recognize these dynamics on some level, but unconscious paradigm influences constrain too many researchers from actively integrating this presumption into their actual analytic practice. The pernicious practice of saying, "the test is

reliable," creates a language that unconsciously predisposes researchers against acting on a conscious realization that tests themselves are not reliable (Thompson, 1994c). Reinhardt (1996) provides an excellent relevant review of reliability coefficients, and the factors that impact score reliability.

As Rowley (1976, p. 53, emphasis added) argued, "It needs to be established that an instrument itself is neither reliable nor unreliable.... A single instrument can produce *scores* which are reliable, and other *scores* which are unreliable." Similarly, Crocker and Algina (1986, p. 144, emphasis added) argued that, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the *scores* on a test for a *particular* group of examinees."

In another widely respected text, Gronlund and Linn (1990, p. 78, emphasis in original) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself.... Thus, it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument."

And Eason (1991, p. 84, emphasis added) argued that:

> Though some practitioners of the classical measurement paradigm [incorrectly] speak of reliability as a characteristic of tests, in fact reliability is a characteristic of *data*, albeit data generated on a given measure administered with a

given protocol to given subjects on given occasions.

The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of "the reliability of the test" without considering to whom the test was administered, or other facets of each individual measurement protocol. Reliability is driven by variance--typically, greater score variance leads to greater score reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogenous or to more homogeneous sets of subjects, will yield scores with differing reliability. As Dawis (1987, p. 486) observed, "[b]ecause reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point."

Our shorthand ways of speaking (e.g., language saying "the test is reliable") can itself cause confusion and lead to bad practice. As Pedhazur and Schmelkin (1991, p. 82, emphasis in original) observed, "Statements about *the* reliability of a measure are... inappropriate and potentially misleading." These telegraphic ways of speaking are not inherently problematic, but they often later become so when we come unconsciously to ascribe literal truth to our shorthand, rather than recognizing that our jargon is merely telegraphic and is <u>not</u> literally true. As noted elsewhere:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious

outcome, sloppy thinking and sloppy practice. (Thompson, 1992c, p. 436)

Implications for Practice

These views suggest at least three implications for research practice. These practices are, unfortunately, not yet normative within the social sciences.

Language Use. One fairly straightforward recommendation is that researchers should not use language saying that, "the test is reliable [or valid]," or that, "the reliability [or validity] of the test was .xx." Because on its face this language is inaccurate, and asserts untruth, it seems imprudent to use such language in scholarly discourse. The editorial policies of at least one journal commend better, correct practices:

> Based on these considerations, use of wording such as "the reliability of the test" or "the validity of the test" will not be considered acceptable in the journal. Instead, authors should use language such as, "the scores in our study had a classical theory test-retest reliability coefficient of X," or "based on generalizability theory analysis, the scores in our study had a phi coefficient of X." Use of technically correct language will hopefully reinforce better practice. (Thompson, 1994c, p. 841)

Coefficient Reporting. Researchers also ought to routinely report the reliability coefficients for their own data. Many do not do so now, because they act under the pernicious misconception that tests are reliable, and are therefore invariant across

administrations.

But it is sloppy practice to not calculate, report, and interpret the reliability of one's own scores for one's own data. As Pedhazur and Schmelkin (1991, p. 86, emphasis in original) argued:

> Researchers who bother at all to report reliability estimates for the instruments they use (many do not) frequently report only reliability estimates contained in the manuals of the instruments or estimates reported by other researchers. Such information may be useful for comparative purposes, but it is imperative to recognize that the *relevant reliability estimate is the one obtained for the sample used in the [present] study under consideration*.

Unhappily, empirical studies indicate that such reports are infrequent (Meier & Davis, 1990; Willson, 1980) in most journals, although there are exceptions (Thompson & Snyder, in press).

In her important paper proposing "reliability generalization" methods to characterize (a) the mean and (b) the standard deviation of score reliabilities for a given instrument across studies, and to explore (c) the sources of variability in score reliabilities, Vacha-Haase noted a benefit from the routine reporting of score reliability even in substantive studies:

> Furthermore, if authors of empirical studies routinely report reliability coefficients, even in substantive studies, the field will cumulate more

> evidence regarding the psychometric integrity of scores. Such practices would provide more fodder for reliability generalization analyses focusing upon the differential influences of various sources of measurement error. (Vacha-Haase, 1998a, p. 14)

Interpret Results in a Reliability Context. Effect sizes can and should be computed in all studies; Kirk (1996) and Snyder and Lawson (1993) provide excellent reviews of the many options. When and if these effects are deemed (a) noteworthy in magnitude and (b) replicable, then (and only then) these effect sizes should also be interpreted.

Score reliability is one of the several study features that impact detected effects. Score measurement errors always attenuate computed effects to some degree (Schneider & Darcy, 1984). This attenuation ought to be considered when interpreting reported effects. As I have noted elsewhere,

> The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. For example, we may conduct studies that could not possibly yield noteworthy effect sizes, given that score reliability inherently attenuates effect sizes. Or we may not accurately interpret the effect sizes in our studies if we do not consider the reliability of the scores we are actually analyzing. (Thompson, 1994c, p. 840)

Error #5: Incorrectly Interpreting Statistical Significance; Failing to Report Effect Sizes

As Pedhazur and Schmelkin (1991) noted, "probably very few methodological issues have generated as much controversy" (p. 198) as have the use and interpretation of statistical significance tests. These tests have proven surprisingly resistant to repeated efforts "to exorcise the null hypothesis" (Cronbach, 1975, p. 124). Especially noteworthy among the historical efforts to accomplish the exorcism have been works by Rozeboom (1960), Morrison and Henkel (1970), Carver (1978), Meehl (1978), Shaver (1985), and Oakes (1986).

More recently, a seemingly periodic series of articles on the extraordinary limits of statistical significance tests has been published in the American Psychologist (cf. Cohen, 1990, 1994; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989). The entire Volume 61, Number 4 issue of the Journal of Experimental Education was devoted to these themes. Schmidt's (1996) APA Division 5 presidential address was published as the lead article in the second issue of the inagural volume of the new APA journal, Psycholgical Methods. The lead section (cf. Hunter, 1997) of the January, 1997 issue of Psychological Science was devoted to this controversy. The April, 1998 issue of Educational and Psychological Measurement featured two lengthy reviews (Levin, 1998; Thompson, 1998) of a major text (Harlow, Mulaik & Steiger, 1997) on the controversy. And the APA Task Force on Statistical Inference (Shea, 1996) has now been working for nearly two years on related recommendations for improving practices.

Illustrative condemnations of contemporary statistical testing practices can be noted. For example, Schmidt and Hunter (1997) recently argued that "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" (p. 37). Rozeboom (1997) was equally direct:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism... (p. 335)

But, without much question, two articles by the late Jacob Cohen (1990, 1994) have been the most influential. Roger Kirk (1996) characterized the two American Psychologist articles by Cohen as "classics," and argued that "the one individual most responsible for bringing the shortcomings of hypothesis testing to the attention of behavioral and educational researchers is Jacob Cohen" (p. 747).

This onslaught of criticism has provoked reactive advocacy for statistical tests (cf. Cortina & Dunlap, 1997; Frick, 1996; Greenwald, Gonzalez, Harris & Guthrie, 1996; Hagen, 1997; Robinson & Levin, 1997). Some of these treatments have been thoughtful, but others have been seriously flawed (see Thompson, in press-c, in press-d).

Yet, notwithstanding the long-term availability of these many publications, even today some researchers still do not understand what their statistical significance tests do and do not do.

Empirical studies of researcher perceptions of test results confirm that researchers manifest these misconceptions (cf. Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993). Similarly, content reviews of the most widely-used statistics textbooks show that even our most distinguished methodologists do not have a good grasp on the meaning of statistical significance tests (Carver, 1978).

My own views have been articulated in various locations (e.g., Thompson, 1993, 1994d, 1997a, in press-a, in press-d). I believe that three other essays (Thompson, 1996, 1998, in press-b) are particularly noteworthy. And a short, *public-domain* ERIC Digest I published (Thompson, 1994b) may be very useful as a class handout.

I have never argued that significance tests should be banned, though obviously others have argued that view (cf. Carver, 1978; Schmidt & Hunter, 1997). As an author, I do report (without much excitement) the results of statistical significance tests. As an editor of three journals, I have accepted for publication manuscripts that report these tests.

Common Misconceptions Regarding Statistical Tests

In various locations I have criticized common misconceptions regarding the meaning and value of statistical tests (cf. Thompson, 1996, in press-b). Three of these I now briefly summarize here.

Statistical Significance Does Not Test Result Importance. Put simply, improbable events are not intrinsically interesting. Some highly improbable events, in fact, are completely inconsequential. In his classic hypothetical dialogue between two teachers, Shaver

(1985, p. 58) poignantly illustrated the folly of equating result improbability with result importance:

> Chris: ...I set the level of significance at .05, as my advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurrence like that *surely* must be important.
>
> Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... Well, that must have been a truly important occurrence then?

Even more importantly, since the premises of statistical significance tests do not invoke human values, in valid logical argument statistical results therefore can not under any circumstances contain as part of their conclusions information about result value. As I have noted previously, "If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating p's, and so p's cannot be blithely used to infer the value of research results" (Thompson, 1993, p. 365). Thus, statistical tests cannot reasonably be used as an atavistic escape from responsibility for defending result importance (Thompson, 1993), or to maintain a mantle of feigned objectivity (Thompson, in press-b).

Statistical Significance Does Not Test Result Replicability. Social scientists seek to identify relationships that recur under stated conditions. Discovering analogs of cold fusion will make us extremely popular (free drinks, much dancing, etc.) at our next scholarly meeting, but we will eternally thereafter be shunned (no one will accept the drinks we attempt to buy for them, so much for the dancing, etc.) at all future conferences, once our results are discovered to be non-replicable. [So, only report non-replicable results at your last conference, immediately prior to retirement.]

Too many researchers, consciously or unconsciously, incorrectly assume that the p values calculated in statistical significance tests evaluate the probability that results will replicate (Carver, 1978, 1993). But statistical tests do not evaluate the probability that the sample statistics occur in the population as parameters (Cohen, 1994).

Instead, "$p_{CALCULATED}$ is the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true" (Thompson, 1996, p. 27). Obviously, knowing the probability of the sample is less interesting than knowing the probability of the population. Knowing the probability of population parameters would bear upon result replicability, since we would then know something about the population from which future researchers would also draw their samples.

But as Shaver (1993) argued so emphatically:

> [A] test of statistical significance is not an indication of the probability that a result would be

> obtained upon replication of the study.... Carver's (1978) treatment should have dealt a death blow to this fallacy.... (p. 304)

And so Cohen (1994) concluded that the statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997).

Statistical Significance Does Not Solely Evaluate Effect Magnitude. Because various study features (including score reliability) impact calculated *p* values, $p_{CALCULATED}$ cannot be used as a satisfactory index of study effect size. As I have noted elsewhere,

> The calculated *p* values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because *p* values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $p_{CALCULATED}$. (Thompson, in press-b)

The recent fourth edition of the American Psychological Association style manual (APA, 1994) explicitly acknowledges that *p* values are not acceptable indices of effect:

> Neither of the two types of probability values [statistical significance tests] reflects the

> importance or magnitude of an effect because both depend on sample size... You are [therefore] *encouraged* to provide effect-size information. (APA, 1994, p. 18, emphasis added)

Recommended Improvements in Statistical Testing Practices

In various locations (cf. Thompson, 1996, in press-b) I have advocated certain changed practices as regards the use of statistical tests. Five such suggested changes are now summarized here.

Effect Sizes Should Be Reported for *All* Tested Effects. The single most important potential improvement in analytic practice would be the regular and routine reporting of effect sizes in all studies. As noted previously, such reports are at least "encouraged" by the new APA (1994, p. 18) style manual.

However, empirical studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "*encouraging*" effect size reporting (APA, 1994) has not appreciably affected actual reporting practices (e.g., Kirk, 1996; Snyder & Thompson, in press; Thompson & Snyder, 1997, in press; Vacha-Haase & Nilson, in press). An on-going series of additional empirical studies of reporting practices has yielded similar results for yet more journals (Lance & Vacha-Haase, 1998; Ness & Vacha-Haase, 1998; Nillson & Vacha-Haase, 1998; Reetz & Vacha-Haase, 1998).

Effect sizes are important to report for at least two reasons. First, when these effects are noteworthy, these indices inform judgment regarding the practical or substantive significance of

results (cf. Kirk, 1996). Second, reporting all effect sizes (even non-statistically significant effects, though some might not interpret them) facilitates the meta-analytic integration of findings across a given literature.

There are many effect sizes (e.g., "uncorrected," "corrected," standardized differences) that can be computed (cf. Kirk, 1996; Snyder & Lawson, 1993). In my view (Thompson, in press-b), arguments can be made that certain indices should be preferred over others. But the important point is that, as regards effect size reporting, it is generally better to report *anything* as against nothing, which is the effect size that most researchers currently report.

Of course, an effect size is no more magical than is statistical significance testing, for the two reasons noted by Zwick (1997). First, because human values are also not part of the calculation of an effect size, any more than values are part of the calculation of p, "largeness of effect does not guarantee practical importance any more than statistical significance does" (p. 4).

Second, some researchers have too rigidly adopted Cohen's (1988) definitions of small, medium and large effects, just as some researchers too rigidity adopted "α=.05" as their gold standard. Cohen (1988) only intended these as impressionistic characterizations of result typicality across a diverse literature. However, some empirical studies do suggest that the characterization is reasonably accurate (Glass, 1979; Olejnik, 1984), at least as regards a literature historically built with a bias against statistically non-significant results (Rosenthal,

1979).

In my view, editorial requirements (Vacha-Haase, 1998b) will ultimately be required to move the field to change analytic and reporting practices. Fortunately, editorial policies at some journals now *require* authors to report and interpret effect sizes. For example, the author guidelines of the Journal of Experimental Education indicate that "authors are required to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (Heldref Foundation, 1997, pp. 95-96, emphasis added). I believe the EPM author guidelines are equally informed:

> We will go further [than mere encouragement]. Authors reporting statistical significance will be required to both report and interpret effect sizes. However, these effect sizes may be of various forms, including standardized differences, or uncorrected (e.g., r^2, R^2, eta^2) or corrected (e.g., adjusted R^2, $omega^2$) variance-accounted-for statistics. (Thompson, 1994c, p. 845, emphasis in original)

It is particularly noteworthy that editorial policies even at one APA journal now indicate that:

> If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is

> a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

Researchers Should More Frequently Employ Non-Nill Nulls. An important but overlooked (see Hagen, 1997; Thompson, in press-c) element of Cohen's (1994) classic article involved his striking criticism of the routine use of "nil" null hypotheses. Cohen (1994) defined a "nil" null hypothesis as a null specifying no differences (e.g., SD_1-SD_2 = 0) or zero correlations (e.g., R^2=0).

Some researchers employ nil nulls because statistical theory does not easily accommodate the testing of some non-nil nulls. But in other cases researchers employ nil nulls because these nulls have been unconsciously accepted as traditional, because these nulls can be mindlessly formulated without consulting previous literature, or because most computer software defaults to tests of nil nulls (Thompson, 1998, in press-b, in press-c).

Unfortunately, when a statistical significance test presumes a nil null is true in the population, an untruth is posited. As Meehl (1978, p. 822) noted, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Similarly, Hays (1981, p. 293) pointed out that "[t]here is surely nothing on earth that is completely independent of anything else [in the population]. The strength of association may approach zero, but it should seldom or never be exactly zero."

Highly respected statistician Roger Kirk (1996) put the point succinctly in his important recent article:

> Because the null hypothesis is *always* false, a decision to reject it simply indicates that the research design had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect. It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers to focus on controlling the Type I error that cannot occur because *all* null hypotheses are false. (p. 747, emphasis added)

And a $p_{CALCULATED}$ value computed on the foundation of a false premise is inherently of somewhat limited utility.

There is a very important implication of the realization that the nil null is untrue in the population. As Hays (1981, p. 293) emphasized, because the nil null is untrue in the population, sample statistics should reflect some difference or some effect, and thus "virtually any study can be made to show significant results if one uses enough subjects." This means that

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. (Thompson, 1992c, p. 436)

Statistical significance would be considerably more informative if researchers reviewed relevant previous research, and then

constructed hypotheses that incorporated previous results.

Measurement Results Should be Tested with Non-Nil Nulls. There is growing recognition that some uses of statistical tests in measurement studies, as regards reliability or validity coefficients or construct validity tests of means, can be particularly misguided. For example, Abelson (1997) commented on statistical tests of measurement study results using nil null hypotheses:

> And when a reliability coefficient is declared to be nonzero, that is the ultimate in stupefyingly vacuous information. What we really want to know is whether an estimated reliability is .50'ish or .80'ish. (Abelson, 1997, p. 121)

Fortunately, the author guidelines of some journals have become more enlightened as regards such practices:

> Statistical tests of such coefficients in a measurement context make little sense. Either statistical significance tests using the [nil] null hypothesis of zero magnitude should be by-passed, or meaningful null hypotheses should be employed. (Thompson, 1994c, p. 844)

Researchers Should Provide Some Warrant That Results Are Replicable. Because evidence of result replicability is important (if we take science to be the business of cumulating knowledge across studies), because statistical significance tests do not evaluate result replicability (Cohen, 1994; Thompson, 1996, 1997b), other methods must and should be used for this purpose. It has been

suggested that

> As more researchers finally realize that statistical significance tests do not test the population, and therefore do not test replicability, researchers will increasingly emphasize evidence that instead is relevant to the issue of result replicability. (Vacha-Haase & Thompson, in press)

Many warrants are available, and in fact a single study might present several such warrants.

The most persuasive, and perhaps the only conclusive, evidence for result replicability is to actually replicate the study. And replication studies are important, and probably are somewhat undervalued in the social sciences (Robinson & Levin, 1997). However, many researchers (especially doctoral students working on dissertations and junior faculty seeking tenure) find themselves unable to replicate every study.

One potential warrant for replicability would involve prospectively formulating null hypotheses by reflectively consulting the effect sizes reported in previous related studies, and by prospectively interpreting study effects in the context of specific previous findings. In effect, virtually any study might be conducted and interpreted as a partial replication of previous inquiry. Another alternative warrant involves empirical investigation of replicability by conducting what I have termed (cf. Thompson, 1996) "internal" replicability analyses.

"Internal" replicability analyses empirically use the sample in hand to combine the participants in different ways to estimate

how much the idiosyncracies of individuality within the sample have compromised generalizability. The major "internal" empirical replicability analyses are cross-validation, the jackknife, and the bootstrap (Diaconis & Efron, 1983); the logics are reviewed in more detail elsewhere (cf. Thompson, 1993, 1994d). "Internal" evidence for replicability is never as good as an actual replication (Robinson & Levin, 1997; Thompson, 1997a), but is certainly better than incorrectly presuming that statistical significance assures result replicability.

However, it must be emphasized that the inferential and the descriptive uses of these logics should not be confused (Thompson, 1993). For example, the *inferential* use of the bootstrap involves using the bootstrap to estimate a sampling distribution when the sampling distribution is not known or assumptions for the use of a known sampling distribution cannot be met (i.e., to conduct a different form of statistical significance test). The *descriptive* use of the bootstrap looks primarily at the variability in effect sizes or other parameter estimates across many different combinations of the participants. The software to conduct "internal" bootstrap analyses for statistics commonly used in the social sciences (cf. Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985) is already widely available (e.g., Lunneborg (1987) for univariate applications, and Thompson (1988b, 1992a, 1995a) for multivariate applications).

Improved Language Use. In Thompson (1996), I suggested that when the null hypothesis is rejected, "such results ought to always be described as 'statistically significant,' and should never be

described only as 'significant'" (pp. 28-29). My argument (Thompson, 1996, 1997a; but see Robinson & Levin, 1997) has been that the common meaning of "significant" has nothing to do with the statistical use of this term, and that the use of the complete phrase might help at least some in conveying that this technical phrase has nothing to do with result importance.

Carver (1993) eloquently made the same argument:

> When trying to emulate the best principles of science, it seems important to say what we mean and to mean what we say. Even though many readers of scientific journals know that the word *significant* is supposed to mean *statistically significant* when it is used in this context, many readers do *not* know this. Why be unnecessarily confusing when clarity should be most important? (p. 288, emphasis in original)

Summary

After presenting a general linear model as a framework for discussion, the present paper reviewed five methodology errors that occur in educational research: (a) the use of *stepwise* methods; (b) the failure to consider in result interpretation the *context specificity* of analytic weights (e.g., regression beta weights, factor pattern coefficients, discriminant function coefficients, canonical function coefficients) that are part of all parametric quantitative analyses; (c) the failure to interpret *both weights and structure coefficients* as part of result interpretation; (d) the failure to recognize that *reliability* is a characteristic of

scores, and not of tests; and (e) the incorrect interpretation of *statistical significance* and the related failure to report and interpret the *effect sizes* present in all quantitative analyses. In several cases small heuristic discriminant analysis data sets were presented to make more concrete and accessible the discussion of each of these five methodology errors.

However, of the various arenas for improvement, the one where I believe the most progress could be realized involves the use of statistical significance tests and the reporting of effect sizes. Yet this is where the most resistance has seemingly occurred. For example, Schmidt and Hunter (1997) recently argued that "logic-based arguments seem to have had only a limited impact... [perhaps due to] the virtual brainwashing in significance testing that all of us have undergone" (pp. 38-39). They also spoke of a "psychology of addiction to significance testing" (Schmidt & Hunter, 1997, p. 49).

Journal editor Loftus (1994), like others, has lamented that repeated publications of

> these concerns never seem to attract much attention (much less impel action). They are carefully crafted and put forth for consideration, only to just kind of dissolve away in the vast acid bath of our existing methodological orthodoxy. (p. 1)

Another editor commented: "p values are like mosquitos" that apparently "have an evolutionary niche somewhere and [unfortunately] no amount of scratching, swatting or spraying will dislodge them" (Campbell, 1982, p. 698).

Similar comments have been made by non-editors. For example, Falk and Greenbaum (1995) noted that "A massive educational effort is required to... extinguish the mindless use of a procedure that dies hard" (p. 94). And Harris (1991) observed, "it is surprising that the dragon will not stay dead" (p. 375).

Fortunately, some slow, glacial progress in the incremental movement of the field was reflected in the APA (1994, p. 18) style manual "encouraging" the reporting of effect sizes. But enlightened editorial policies (e.g., Heldref Foundation, 1997; Murphy, 1997; Thompson, 1994c) now provide the strongest basis for cautious optimism.

References

Abelson, R.P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 117-141). Mahwah, NJ: Erlbaum.

Aiken, L.S., West, S.G., Sechrest, L., Reno, R.R., with Roediger, H.L., Scarr, S., Kazdin, A.E., & Sherman, S.J. (1990). The training in statistics, methodology, and measurement in psychology. American Psychologist, 45, 721-734.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Bagozzi, R.P. (1981). Canonical correlation analysis as a special case of a structural relations model. Multivariate Behavioral Research, 16, 437-454.

Borgen, F.H., & Seling, M.J. (1978). Uses of discriminant analysis following MANOVA: Multivariate statistics for multivariate purposes. Journal of Applied Psychology, 63(6), 689-697.

Campbell, N. (1982). Editorial: Some remarks from the outgoing editor. Journal of Applied Psychology, 67, 691-700.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cooley, W.W., & Lohnes, P.R. (1971). Multivariate data analysis. New York: John Wiley & Sons.

Cortina, J.M., & Dunlap, W.P. (1997). Logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Dawis, R.V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

---

* Cited empirical studies of methodological practice are designated with asterisks.

Duncan, O.D. (1975). Introduction to structural equation models. New York: Academic Press.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.

*Emmons, N.J., Stallings, W.M., & Layne, B.H. (1990, April). Statistical methods used in *American Educational Research Journal*, *Journal of Educational Psychology*, and *Sociology of Education* from 1972 through 1987. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 319 797)

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, 5(1), 75-98.

Fan, X. (1996). Canonical correlation analysis as a general analytic model. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 71-94). Greenwich, CT: JAI Press.

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? Structural Equation Modeling, 4, 65-79.

Fish, L.J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Gage, N.L. (1985). Hard gains in the soft sciences: The case of pedagogy. Bloomington, IN: Phi Delta Kappa Center on Evaluation, Development, and Research.

Gall, M.D., Borg, W.R., & Gall, J.P. (1996). Educational research: An introduction (6th ed.). White Plains, NY: Longman.

*Glass, G.V (1979). Policy for the unpredictable (uncertainty research and policy). Educational Researcher, 8(9), 12-14.

Goodwin, L.D., & Goodwin, W.L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5-11.

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect size and *p*-values: What should be reported and what should be replicated? Psychophysiology, 33(2), 175-183.

Grimm, L.G., & Yarnold, P.R. (Eds.). (1995). Reading and understanding multivariate statistics. Washington, DC: American Psychological Association.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.

*Hall, B.W., Ward, A.W., & Comer, C.B. (1988). Published educational research: An empirical study of its quality. Journal of Educational Research, 81, 182-189.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.). (1997). What if there were no significance tests?. Mahwah, NJ: Erlbaum.

Harris, M.J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. Theory & Psychology, 1, 375-382.

Harris, R.J. (1989). A canonical cautionary. Multivariate Behavioral Research, 24, 17-39.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Heldref Foundation. (1997). Guidelines for contributors. Journal of Experimental Education, 65, 95-96.

Henard, D.H. (1998, January). Suppressor variable effects: Toward understanding an elusive data dynamic. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED forthcoming)

Holzinger, K. L. & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution (No. 48). Chicago: University of Chicago.

Horst, P. (1966). Psychological measurement and prediction. Belmont, CA: Wadsworth.

Huberty, C.J (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Huberty, C.J (1994). Applied discriminant analysis. New York: Wiley and Sons.

Huberty, C.J, & Barton, R. (1989). An introduction to discriminant analysis. Measurement and Evaluation in Counseling and Development, 22, 158-168.

Huberty, C.J, & Lowman, L.L. (1997). Discriminant analysis via statistical packages. Educational and Psychological Measurement, 57, 759-784.

Huberty, C.J, & Wisenbaker, J. (1992). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2, pp. 169-208). Greenwich, CT: JAI Press.

Huebner, E. S. (1991). Correlates of life satisfaction in children. School Psychology Quarterly, 6, 103-111.

Huebner, E. S. (1992). Burnout among school psychologists: An exploratory investigation into its nature, extent, and correlates. School Psychology Quarterly, 7, 129-136.

Humphries-Wadsworth, T.M. (1998, April). Features of published analyses of canonical results. Paper presented at the annual meeting of the American Educational Research Association, San Diego. (ERIC Document Reproduction Service No. ED forthcoming)

Hunter, J.E. (1997). Needed: A ban on the significance test. Psychological Science, 8(1), 3-7.

Jöreskog, K.G., & Sörbom, D. (1989). LISREL 7: A guide to the

program and applications (2nd ed.). Chicago: SPSS.

Jorgenson, C. B., Jorgenson, D. E., Gillis, M. K., & McCall, C. M. (1993). Validation of a screening instrument for young children with teacher assessment of school performance. School Psychology Quarterly, 8, 125-139.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

*Kirk, R. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

*Lance, T., & Vacha-Haase, T. (1998, August). *The Counseling Psychologist*: Trends and usages of statistical significance testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Levin, J.R. (1998). To test or not to test H_0? Educational and Psychological Measurement, 58, 311-331.

Levine, M. S. (1977). Canonical analysis and factor comparison. Newbury Park, CA: Sage.

Loftus, G.R. (1994, August). Why psychology will never be a real science until we change the way we analyze data. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Lunneborg, C.E. (1987). Bootstrap applications for the behavioral sciences. Seattle: University of Washington.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

*Meier, S.T., & Davis, S.R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.

Meredith, W. (1964). Canonical correlations with fallible data. Psychometrika, 29, 55-65.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Murphy, K.R. (1997). Editorial. Journal of Applied Psychology, 82, 3-5.

*Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

*Ness, C., & Vacha-Haase, T. (1998, August). Statistical significance reporting: Current trends and usages within *Professional Psychology: Research and Practice*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

*Nillson, J., & Vacha-Haase, T. (1998, August). A review of statistical significance reporting in the *Journal of Counseling Psychology*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

*Oakes, M. (1986). Statistical inference: A commentary for the

social and behavioral sciences. New York: Wiley.

*Olejnik, S.F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

*Reetz, D., & Vacha-Haase, T. (1998, August). Trends and usages of statistical significance testing in adult development and aging research: A review of *Psychology and Aging*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher, 26(5), 21-26.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

*Rosenthal, R. & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. Journal og Psychology, 55, 33-38.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1(2), 115-129.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8, 573-582.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42(49), A12, A16.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), (1991). Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 99-105). Greenwich, CT: JAI Press.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Snyder, P.A., & Thompson, B. (in press). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. School Psychology Quarterly.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park, CA: Sage.

Thompson, B. (1985). Alternate methods for analyzing data from experiments. Journal of Experimental Education, 54, 50-55.

Thompson, B. (1988a, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1988b). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. Educational and Psychological Measurement, 48, 681-686.

Thompson, B. (1989). Why won't stepwise methods die?. Measurement and Evaluation in Counseling and Development, 21(4), 146-148.

Thompson, B. (1990). ALPHAMAX: A program that maximizes coefficient alpha by selective item deletion. Educational and Psychological Measurement, 50, 585-589.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1992a). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids. Educational and Psychological Measurement, 52, 905-911.

Thompson, B. (1992b). Misuse of ANCOVA and related "statistical control" procedures. Reading Psychology, 13, iii-xviii.

Thompson, B. (1992c). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1994a, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1994b). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). Measurement Update, 4(1), 5-6. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B. (1994c). Guidelines for authors. Educational and Psychological Measurement, 54(4), 837-847.

Thompson, B. (1994d). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62, 157-176.

Thompson, B. (1994e, February). Why multivariate methods are usually vital in research: Some basic concepts. Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Society for Research in Human Development, Austin, TX. (ERIC Document Reproduction Service No. ED 367 687)

Thompson, B. (1995a). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Educational and Psychological Measurement, 55, 84-94.

Thompson, B. (1995b). Review of *Applied discriminant analysis* by C.J Huberty. Educational and Psychological Measurement, 55, 340-350.

Thompson, B. (1995c). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1997a). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher, 26(5), 29-32.

Thompson, B. (1997b). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. Educational and Psychological Measurement, 57, 5-19.

Thompson, B. (1998). Review of *What if there were no significance tests?* by L. Harlow, S. Mulaik & J. Steiger (Eds.). Educational and Psychological Measurement, 58, 332-344.

Thompson, B. (in press-a). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), Reading and understanding multivariate statistics (Vol. 2). Washington, DC: American Psychological Association.

Thompson, B. (in press-b). If statistical significance tests are broken/misused, what practices should supplement or replace them?. Theory & Psychology.

Thompson, B. (in press-c). In praise of brilliance, where that praise really belongs. American Psychologist.

Thompson, B. (in press-d). Why "encouraging" effect size reporting isn't working: The etiology of researcher resistance to changing practices. Journal of Psychology.

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

Thompson, B., & Daniel, L.G. (1996a). Factor analytic evidence for the construct validity of scores: An historical overview and

some guidelines. Educational and Psychological Measurement, 56, 213-224.

Thompson, B., & Daniel, L.G. (1996b). Seminal readings on reliability and validity: A "hit parade" bibliography. Educational and Psychological Measurement, 56, 741-745.

*Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. Journal of Experimental Education, 66, 75-83.

*Thompson, B., & Snyder, P.A. (in press). Statistical significance and reliability analyses in recent JCD research articles. Journal of Counseling and Development.

Travers, R.M.W. (1983). How research has changed American schools: A history from 1840 to the present. Kalamazoo, MI: Mythos Press.

Tuckman, B.W. (1990). A proposal for improving the quality of published educational research. Educational Researcher, 19(9), 22-24.

Vacha-Haase, T. (1998a). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Vacha-Haase, T. (1998b, August). A review of APA journals' editorial policies regarding statistical significance testing and effect size. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

*Vacha-Haase, T., & Nilsson, J.E. (in press). Statistical significance reporting: Current trends and usages within *MECD*. Measurement and Evaluation in Counseling and Development.

Vacha-Haase, T., & Thompson, B. (in press). Further comments on statistical significance tests. Measurement and Evaluation in Counseling and Development.

*Vockell, E.L., & Asher, W. (1974). Perceptions of document quality and use by educational decision makers and researchers. American Educational Research Journal, 11, 249-258.

*Wandt, E. (1967). An evaluation of educational research published in journals (Report of the Committee on Evaluation of Research). Washington, DC: American Educational Research Association.

*Ward, A.W., Hall, B.W., & Schramm, C.E. (1975). Evaluation of published educational research: A national survey. American Educational Research Journal, 12, 109-128.

*Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

*Zuckerman, M., Hodgins, H.S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. Psychological Science, 4, 49-53.

Zwick, R. (1997, March). Would the abolition of significance testing lead to better science? Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Table 1
Correlation Coefficients for Selected
Holzinger and Swineford (1939) Data Used to Illustrate That
SEM is the Most General Case of the General Linear Model

	T6	T7	T2	T4	T20	T21	T22
T6	1.0000	.7332	.1529	.1586	.3440	.3206	.4476
T7	.7332	1.0000	.1394	.0772	.3367	.3020	.4698
T2	.1529	.1394	1.0000	.3398	.2812	.2433	.2812
T4	.1586	.0772	.3398	1.0000	.3243	.3310	.3062
T20	.3440	.3367	.2812	.3243	1.0000	.3899	.3947
T21	.3206	.3020	.2433	.3310	.3899	1.0000	.3767
T22	.4476	.4698	.2812	.3062	.3947	.3767	1.0000

Note. The variable labels for these seven variables are:
T6 PARAGRAPH COMPREHENSION TEST
T7 SENTENCE COMPLETION TEST
T2 CUBES, SIMPLIFICATION OF BRIGHAM'S SPATIAL RELATIONS TEST
T4 LOZENGES FROM THORNDIKE--SHAPES FLIPPED OVER THEN IDENTIFY TARGET
T20 DEDUCTIVE MATH ABILITY
T21 MATH NUMBER PUZZLES
T22 MATH WORD PROBLEM REASONING

Table 2
Standardized *Canonical Function Coefficients* for the Table 1 Data
Derived Using the Appendix A SPSS/LISREL Program to Illustrate That
SEM is the Most General Case of the General Linear Model

Standardized canonical coefficients for DEPENDENT variables

Variable	1	2
T6	.44962	-1.40007
T7	.62246	1.33225

Standardized canonical coefficients for COVARIATES

COVARIATE	1	2
T2	-.01468	.06704
T4	-.20012	-1.00653
T20	.34100	-.02762
T21	.26772	-.17401
T22	.73104	.35974

Table 3
LISREL *"Gamma" Coefficients* for the Table 1 Data
Derived Using the Appendix A SPSS/LISREL Program to Illustrate That
SEM is the Most General Case of the General Linear Model

GAMMA					
	T6	T7			
ETA 1	0.44957	0.62250			
GAMMA					
	T2	T4	T20	T21	T22
ETA 1	-0.01468	-0.20014	0.34100	0.26772	0.73104
GAMMA					
	T6	T7			
ETA 1	0.44956	0.62251			
ETA 2	1.40013	-1.33228			
GAMMA					
	T2	T4	T20	T21	T22
ETA 1	-0.01469	-0.20014	0.34101	0.26771	0.73104
ETA 2	-0.06706	1.00653	0.02762	0.17402	-0.35972

Note. The LISREL coefficients for the "gamma" matrix exactly match (within rounding error) the canonical function coefficients presented previously. The only exception is that all the signs for the SEM second canonical function coefficients must be "reflected." "Reflecting" a function (changing *all* the signs on a given function, factor, or equation) is *always* permissible, because the scaling of psychological constructs is arbitrary. Thus, the SEM and the canonical analysis derived the same results.

Table 4
The Confusing Language of Statistics
(Intentionally Designed to Confuse the Graduate Students)

Analysis	Standardized Weights*	Weight System	Snythetic/ Latent Variable(s)
Multiple Regression	β	"equation"	Yhat ($\hat{Y}$)
Factor Analysis	pattern coefficients	"factor"	factor scores
Descriptive Discriminant Analysis	standardized function coefficients	"function" -or- "rule"	discriminant function scores
Canonical Correlation Analysis	standardized function coefficients	"function"	canonical function scores

*Of course, the term, "standardized weight", is an obvious oxymoron. A given weight is a *constant* applied to all the scores of all the cases/people on the observed/manifest/ measured variable, and therefore *cannot* be standardized. Instead, the weighting constant is applied to the measured variable in its standardized form, i.e., we should say "weight for the standardized measured variables" rather than "standardized weight".

Table 5
Holzinger and Swineford *Data* to Show That
More Predictors May Actually *Hurt* Classification Accuracy

Seq	ID	GRADE	T13	T17	T22	T16
1	2	7	285	12	21	100
2	3	7	159	1	18	95
3	9	7	265	18	18	105
4	14	7	211	8	22	103
5	16	7	211	5	34	102
6	18	7	189	13	16	100
7	20	7	207	3	47	107
8	22	7	194	8	19	96
9	25	7	244	6	20	99
10	28	7	163	12	24	106
11	30	7	310	10	20	101
12	34	7	121	3	18	92
13	44	7	167	11	22	112
14	46	7	100	4	25	58
15	47	7	240	6	20	103
16	50	7	226	4	39	109
17	51	7	196	8	18	96
18	52	7	218	7	18	92
19	58	7	151	15	25	102
20	66	7	142	3	13	95
21	68	7	172	10	32	110
22	71	7	181	9	27	107
23	74	7	153	15	21	99
24	75	7	141	14	19	107
25	76	7	195	10	19	103
26	78	7	186	7	30	109
27	79	7	215	10	15	103
28	81	7	165	11	22	108
29	83	7	233	2	28	100
30	85	7	203	8	24	103
31	202	7	195	9	22	106
32	203	7	228	1	43	101
33	205	7	160	9	35	99
34	208	7	333	16	45	118
35	213	7	154	3	19	106
36	225	7	236	21	29	116
37	226	7	219	6	23	104
38	230	7	189	1	7	99
39	232	7	143	2	27	94
40	235	7	162	3	16	100
41	236	7	205	6	27	101
42	239	7	112	3	18	90
43	244	7	137	0	24	105
44	245	7	214	4	26	100
45	250	7	120	3	28	112
46	252	7	165	1	10	101

47	253	7	137	1	15	89
48	256	7	214	4	28	97
49	257	7	223	5	23	106
50	263	7	205	5	35	103
51	264	7	180	6	36	97
52	268	7	130	3	14	103
53	269	7	220	4	31	113
54	277	7	149	1	21	96
55	86	8	207	19	37	112
56	88	8	217	24	20	106
57	89	8	191	10	27	109
58	90	8	208	9	17	98
59	106	8	260	17	41	104
60	112	8	148	11	34	105
61	118	8	271	11	34	113
62	120	8	175	10	24	111
63	126	8	180	11	21	96
64	131	8	247	20	26	101
65	132	8	119	2	28	91
66	133	8	234	14	44	113
67	134	8	172	23	26	99
68	137	8	177	11	25	93
69	139	8	208	18	34	107
70	140	8	227	9	13	108
71	143	8	259	16	23	107
72	148	8	196	7	39	96
73	150	8	248	17	32	110
74	151	8	255	26	34	112
75	153	8	206	11	16	105
76	155	8	238	16	49	102
77	158	8	227	18	15	101
78	160	8	197	6	25	100
79	165	8	195	9	29	91
80	282	8	241	1	27	115
81	283	8	230	4	26	103
82	284	8	200	11	8	108
83	285	8	246	16	33	109
84	287	8	227	11	48	109
85	288	8	168	11	28	104
86	289	8	224	13	43	104
87	290	8	189	7	38	110
88	297	8	199	8	30	108
89	298	8	249	15	50	119
90	299	8	212	7	29	102
91	304	8	210	5	27	104
92	311	8	198	7	34	107
93	312	8	237	6	18	108
94	313	8	206	15	50	107
95	315	8	215	5	27	101
96	317	8	183	9	18	113
97	318	8	187	8	35	109
98	322	8	220	7	26	109
99	323	8	178	8	27	103

100	324	8	150	6	8	102
101	329	8	235	6	18	101
102	338	8	206	26	37	113
103	341	8	174	7	46	105
104	342	8	162	9	29	96
105	343	8	228	1	39	104
106	345	8	204	7	25	112
107	351	8	186	25	39	109

Note. The variable labels are:
T13 SPEEDED DISCRIM STRAIGHT AND CURVED CAPS
T17 MEMORY OF OBJECT-NUMBER ASSOCIATION TARGETS
T22 MATH WORD PROBLEM REASONING
T16 MEMORY OF TARGET SHAPES

Table 6
Holzinger and Swineford *Results* to Show That
More Predictors May Actually *Hurt* Classification Accuracy
--*LDF* and *LCF* Score Classification Tables--

GRADE by LDFCL3 *LDF* classification **3 predictors**

Count	7	8	Row Total
GRADE 7	**40**	14	54
			50.5
8	22	**31**	53
			49.5
Column Total	62	45	107
	57.9	42.1	100.0

GRADE by LDFCL4 *LDF* classification **4 predictors**

Count	7	8	Row Total
GRADE 7	**38**	16	54
			50.5
8	23	**30**	53
			49.5
Column Total	61	46	107
	57.0	43.0	100.0

GRADE by LCFCL3 *LCF* classification **3 predictors**

Count	7	8	Row Total
GRADE			
7	**40**	14	54
			50.5
8	22	**31**	53
			49.5
Column Total	62	45	107
	57.9	42.1	100.0

GRADE by LCFCL4 *LCF* classification **4 predictors**

Count	7	8	Row Total
GRADE			
7	**38**	16	54
			50.5
8	23	**30**	53
			49.5
Column Total	61	46	107
	57.0	43.0	100.0

Table 7
Holzinger and Swineford *Results* to Show That
More Predictors May Actually *Hurt* Classification Accuracy
--Both LDF and LCF Actual Classifications--

Seq	ID	GRADE		LDFCL3	LDFCL4	LCFCL3	LCFCL4
1	2	7		8	8	8	8
2	3	7		7	7	7	7
3	9	7		8	8	8	8
4	14	7		7	7	7	7
5	16	7		7	7	7	7
6	18	7		7	7	7	7
7	20	7		8	8	8	8
8	22	7		7	7	7	7
9	25	7		7	7	7	7
10	28	7		8	8	8	8
11	30	7	+	8	7	8	7
12	34	7		7	7	7	7
13	44	7	-	7	8	7	8
14	46	7		7	7	7	7
15	47	7		7	7	7	7
16	50	7		8	8	8	8
17	51	7		7	7	7	7
18	52	7		7	7	7	7
19	58	7		8	8	8	8
20	66	7		7	7	7	7
21	68	7		8	8	8	8
22	71	7	-	7	8	7	8
23	74	7		8	8	8	8
24	75	7		8	8	8	8
25	76	7		7	7	7	7
26	78	7	-	7	8	7	8
27	79	7		7	7	7	7
28	81	7	-	7	8	7	8
29	83	7		7	7	7	7
30	85	7		7	7	7	7
31	202	7		7	7	7	7
32	203	7		7	7	7	7
33	205	7		8	8	8	8
34	208	7		8	8	8	8
35	213	7		7	7	7	7
36	225	7		8	8	8	8
37	226	7		7	7	7	7
38	230	7		7	7	7	7
39	232	7		7	7	7	7
40	235	7		7	7	7	7
41	236	7		7	7	7	7
42	239	7		7	7	7	7
43	244	7		7	7	7	7
44	245	7		7	7	7	7
45	250	7		7	7	7	7

46	252	7		7	7	7	7
47	253	7		7	7	7	7
48	256	7		7	7	7	7
49	257	7		7	7	7	7
50	263	7		7	7	7	7
51	264	7	+	8	7	8	7
52	268	7		7	7	7	7
53	269	7		7	7	7	7
54	277	7		7	7	7	7
55	86	8		8	8	8	8
56	88	8		8	8	8	8
57	89	8		8	8	8	8
58	90	8		7	7	7	7
59	106	8		8	8	8	8
60	112	8		8	8	8	8
61	118	8		8	8	8	8
62	120	8	+	7	8	7	8
63	126	8		7	7	7	7
64	131	8		8	8	8	8
65	132	8		7	7	7	7
66	133	8		8	8	8	8
67	134	8		8	8	8	8
68	137	8	-	8	7	8	7
69	139	8		8	8	8	8
70	140	8		7	7	7	7
71	143	8		8	8	8	8
72	148	8		8	8	8	8
73	150	8		8	8	8	8
74	151	8		8	8	8	8
75	153	8		7	7	7	7
76	155	8		8	8	8	8
77	158	8		8	8	8	8
78	160	8		7	7	7	7
79	165	8	-	8	7	8	7
80	282	8		7	7	7	7
81	283	8		7	7	7	7
82	284	8		7	7	7	7
83	285	8		8	8	8	8
84	287	8		8	8	8	8
85	288	8		8	8	8	8
86	289	8		8	8	8	8
87	290	8		8	8	8	8
88	297	8		8	8	8	8
89	298	8		8	8	8	8
90	299	8		7	7	7	7
91	304	8		7	7	7	7
92	311	8		8	8	8	8
93	312	8		7	7	7	7
94	313	8		8	8	8	8
95	315	8		7	7	7	7
96	317	8		7	7	7	7
97	318	8		8	8	8	8
98	322	8		7	7	7	7

99	323	8	7	7	7	7
100	324	8	7	7	7	7
101	329	8	7	7	7	7
102	338	8	8	8	8	8
103	341	8	8	8	8	8
104	342	8	7	7	7	7
105	343	8	7	7	7	7
106	345	8	7	7	7	7
107	351	8	8	8	8	8

Note. The variable labels are:
LCFCL3 'LCF classification with 3 preds'
LCFCL4 'LCF classification with 4 preds'
LDFCL3 'LDF classification with 3 preds'
LDFCL4 'LDF classification with 4 preds'

For the present example, for both the 3 and the 4 response variable analyses, the LDF and the LCF scores classified all 107 persons into the same groups. This need not have happened, but will happen as the covariance matrices approach equality across groups.

However, in both the LDF and the LCF analyses, 9 persons were classified differently across these two analyses; these cases are underlined within the table. In both the LDF and the LCF analyses, the use of 4 rather than 3 response variables (a) *correctly* changed the predicted classification of 3 people (denoted with plus signs in the table), (b) *incorrectly* changed the predicted classification of 6 people (denoted with minus signs in the table), thus (c) resulting in a net *worsening* from using more information for prediction as regards 3 persons.

Table 8
Incorrect and Correct Statistical Tests
for Two Steps of Stepwise Analysis
Involving k=3 Groups and n=120 People

Incorrect Step #1
For k=3, p=1
df numerator = n - 1 = 2
df denominator = n - k = 117
lambda = 0.79270

$$F_{exact} = \frac{\frac{1-\Lambda}{\Lambda}}{\frac{n-k}{k-1}}$$

$$\frac{\frac{1-0.79270}{0.79270}}{\frac{117}{2}}$$

$$\frac{\frac{0.20730}{0.79270}}{58.5}$$

$$\frac{0.26151}{58.5}$$

F exact = **15.29841**
p calculated = **.0000012**

Correct Step #1
For k=3, p=50
df numerator = 2 * p = 100
df denominator = 2 (n - p - 2)= 136

$$F_{exact} = \frac{\frac{1-\Lambda^{.5}}{\Lambda^{.5}}}{\frac{n-p-2}{p}}$$

$$\frac{\frac{1-0.79270^{.5}}{0.79270}}{\frac{136}{100}}$$

$$\frac{\frac{1-0.89034}{0.89034}}{\frac{136}{100}}$$

$$\frac{\frac{0.10966}{0.89034}}{1.36}$$

$$\frac{0.12317}{1.36}$$

F exact = **0.16751**
p calculated = **1.00000**

Incorrect Step #2
For k=3, p=2
df numerator = 2 (k - 1) = 4
df denominator = 2 (n - k - 1)= 232
lambda = 0.65540

$$F_{exact} = \frac{\frac{1-\Lambda^{.5}}{\Lambda^{.5}}}{\frac{n-k-1}{k-1}}$$

$$\frac{\frac{1-0.65540^{.5}}{0.65540}}{\frac{232}{4}}$$

$$\frac{\frac{1-0.80957}{0.80957}}{\frac{232}{4}}$$

$$\frac{\frac{0.19043}{0.80957}}{58}$$

$$\frac{0.23523}{58}$$

F exact = **13.64322**
p calculated = **.0000945**

Correct Step #2
For k=3, p=50
df numerator = 2 * p = 100
df denominator = 2 (n - p - 2)= 136

$$F_{exact} = \frac{\frac{1-\Lambda^{.5}}{\Lambda^{.5}}}{\frac{n-p-2}{p}}$$

$$\frac{\frac{1-0.65540^{.5}}{0.65540}}{\frac{136}{100}}$$

$$\frac{\frac{1-0.80957}{0.80957}}{\frac{136}{100}}$$

$$\frac{\frac{0.19043}{0.80957}}{1.36}$$

$$\frac{0.23523}{1.36}$$

F exact = **0.31991**
p calculated = **1.00000**

Note. The formulae for degrees of freedom and F are presented by Tatsuoka (1971, pp. 88-89).

multivar.wk1 3/31/98

Table 9
Heuristic *Data* Illustrating That
Stepwise Methods Do *Not* *Identify the Best* Variable Set

ID	Grp	X1	X2	X3	X4
1	1	30.202	46.146	36.393	44.268
2	1	36.268	44.816	46.370	42.663
3	1	39.381	30.775	32.532	31.966
4	1	32.511	26.201	35.776	40.843
5	1	42.809	39.137	40.845	47.970
6	1	54.841	32.072	32.474	52.689
7	1	32.669	51.460	55.332	40.989
8	1	36.884	45.926	29.255	44.400
9	1	49.781	42.148	43.681	37.719
10	1	51.618	44.373	41.579	48.125
11	1	51.375	43.457	55.160	35.306
12	1	55.102	46.903	44.780	44.669
13	1	33.286	38.660	39.553	32.117
14	1	31.384	41.336	36.259	44.751
15	1	50.000	50.275	61.363	33.207
16	1	39.322	56.273	55.674	34.216
17	1	41.290	47.550	38.913	63.592
18	1	48.098	45.198	38.960	58.692
19	1	61.910	27.474	38.298	46.657
20	1	50.028	51.954	50.832	44.419
21	1	34.585	44.304	36.311	46.899
22	1	57.834	49.899	49.276	50.643
23	1	49.760	29.312	44.098	61.037
24	1	26.010	60.816	58.574	31.081
25	1	23.075	57.059	48.307	40.710
26	1	34.310	44.277	34.315	52.634
27	1	54.714	41.616	51.413	52.284
28	1	60.945	43.890	44.886	40.360
29	1	44.667	52.236	53.525	51.628
30	1	48.442	57.685	57.240	34.324
31	1	38.796	49.830	34.957	45.241
32	1	47.693	43.561	28.529	52.057
33	1	44.497	53.306	41.543	46.079
34	1	55.224	62.785	58.527	32.167
35	1	50.654	26.676	40.851	30.122
36	1	42.632	54.313	49.072	34.758
37	1	50.753	54.410	45.739	59.575
38	1	43.564	42.998	39.366	51.515
39	1	34.850	58.913	64.975	39.955
40	1	50.408	43.214	43.598	59.859
41	2	47.213	37.836	44.151	50.418
42	2	34.168	33.221	29.149	46.838
43	2	58.639	27.033	48.206	52.029
44	2	38.730	49.495	48.813	48.258
45	2	51.596	53.009	51.326	45.759
46	2	62.621	39.735	52.727	71.905

47	2	51.737	37.667	45.013	38.552
48	2	43.922	61.284	55.784	55.129
49	2	42.726	54.703	54.281	37.671
50	2	44.939	48.408	36.004	64.368
51	2	42.050	59.340	61.987	63.012
52	2	37.950	63.446	55.519	35.175
53	2	46.938	56.395	65.436	48.823
54	2	59.976	53.046	51.431	54.273
55	2	59.651	46.707	58.262	48.909
56	2	61.465	36.292	45.301	63.513
57	2	51.051	46.853	51.258	43.695
58	2	40.534	43.357	40.944	50.941
59	2	48.756	53.468	56.950	39.971
60	2	69.683	38.471	49.262	37.572
61	2	46.532	48.917	49.324	62.440
62	2	47.390	33.825	28.706	53.079
63	2	45.617	69.776	56.763	51.743
64	2	56.300	47.684	57.178	51.941
65	2	36.826	69.819	62.206	60.214
66	2	55.413	49.488	48.629	43.843
67	2	52.831	56.210	56.712	45.976
68	2	53.087	46.471	48.024	43.155
69	2	47.221	57.142	52.413	48.072
70	2	54.653	57.012	51.724	48.850
71	2	51.779	65.569	66.259	46.466
72	2	46.009	52.845	48.452	54.614
73	2	52.968	48.023	50.156	50.077
74	2	43.296	45.937	45.162	58.516
75	2	55.779	55.454	59.676	23.961
76	2	55.410	62.863	58.090	48.973
77	2	51.454	57.612	54.929	45.531
78	2	48.538	44.353	49.021	49.085
79	2	62.931	45.867	53.116	54.326
80	2	68.626	47.541	49.993	70.532
81	3	40.113	52.329	50.289	49.856
82	3	63.539	41.711	46.398	59.927
83	3	45.115	61.546	65.551	61.702
84	3	36.029	43.581	38.991	45.273
85	3	51.691	31.516	41.387	55.789
86	3	66.255	59.021	45.930	63.253
87	3	54.119	53.613	57.157	56.673
88	3	49.996	64.174	63.878	61.408
89	3	60.048	59.992	61.433	41.806
90	3	46.350	50.215	59.540	57.780
91	3	49.121	60.275	44.200	69.682
92	3	48.088	68.394	59.637	51.042
93	3	52.787	59.393	61.506	46.042
94	3	44.986	41.866	39.170	43.529
95	3	55.269	68.011	59.191	60.153
96	3	50.261	47.608	44.830	54.833
97	3	56.321	57.470	59.734	51.043
98	3	50.766	49.361	54.050	50.134
99	3	65.540	45.512	58.401	54.444

100	3	47.305	63.725	55.889	44.630
101	3	61.232	52.462	59.623	49.975
102	3	43.688	54.287	54.662	44.419
103	3	74.301	49.445	45.461	64.624
104	3	46.216	55.011	43.794	70.389
105	3	50.882	46.326	42.779	48.925
106	3	48.898	58.229	56.452	60.881
107	3	60.911	60.077	62.039	62.825
108	3	60.918	49.582	43.208	48.960
109	3	49.932	65.463	79.812	53.265
110	3	55.415	61.860	64.733	49.648
111	3	66.505	36.375	41.958	60.718
112	3	59.574	52.291	63.181	60.637
113	3	62.806	42.934	51.890	57.537
114	3	55.761	68.426	60.399	52.615
115	3	73.150	46.255	38.224	77.559
116	3	56.814	60.450	64.211	40.352
117	3	50.092	65.513	44.826	54.327
118	3	65.086	58.518	62.482	48.116
119	3	57.997	66.886	58.486	63.017
120	3	73.867	46.347	70.118	61.087

Table 10
Heuristic *Results* Illustrating That
Stepwise Methods Do *Not* *Identify the Best* Variable Set:
The DDA *Stepwise Results*

AT STEP 1, X1 WAS INCLUDED IN THE ANALYSIS.

		DEGREES OF FREEDOM		SIGNIF.	BETWEEN GROUPS
WILKS' LAMBDA	0.79270	1	2	117.0	
EQUIVALENT F	15.2988		2	117.0	0.0000

* * * \

Table 11
Heuristic *Results* Illustrating That
Stepwise Methods Do *Not Identify the Best* Variable Set:
The DDA *All-Possible-Subsets Results*

X1,X2

DDA	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
	0.6553991	49.223	4	**0.0000**
	0.9992265	0.90148E-01	1	**0.7640**

1-Way MANOVA	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
	.65540	13.64322	4.00	232.00	**.000**
	.99923	.09057	1.00	117.00	**.764**

X1,X3

DDA	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
	0.6961866	42.189	4	0.0000
	0.9988321	0.13614	1	0.7122

1-Way MANOVA	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
	.69619	11.51286	4.00	232.00	.000
	.99883	.13680	1.00	117.00	.712

X1,X4

DDA	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
	0.7081264	40.208	4	0.0000
	0.9991168	0.10294	1	0.7483

1-Way MANOVA	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
	.70813	10.92434	4.00	232.00	.000
	.99912	.10343	1.00	117.00	.748

X2,X3

DDA	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
	0.8094569	24.627	4	0.0001
	0.9913438	1.0128	1	0.3142

1-Way MANOVA	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
	.80946	6.46606	4.00	232.00	.000
	.99134	1.02162	1.00	117.00	.314

X2,X4

DDA	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
	0.6966245	42.116	4	0.0000
	0.9999445	0.64643E-02	1	0.9359

1-Way MANOVA	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
	.69662	11.49101	4.00	232.00	.000
	.99994	.00649	1.00	117.00	.936

X3,X4

DDA	WILKS' LAMBDA	CHI-SQUARED	D.F.	SIGNIFICANCE
	0.6272538	54.336	4	**0.0000**
	0.9973925	0.30417	1	**0.5813**

1-Way MANOVA	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
	.62725	15.23292	4.00	232.00	**.000**
	.99739	.30588	1.00	117.00	**.581**

Note. In addition to illustrating that the stepwise selection of variables X1 and X2 as the first two variables is incorrect, since the lambda value for X3 and X4 is better (**.62725** vs **.65540**), the tabled results also illustrate that DDA and a one-way MANOVA are the same analysis, even though the SPSS programmers made inconsistent choices of test statistics and the number of decimals to report across these two analyses.

Table 12
Heuristic *Data* Illustrating
the *Context Specificity* of GLM Weights

ID/		Response Variable[a]			
Stat.	Grp	X1	X2	X3	X4
1	1	4	3	7	19
2	1	4	4	4	17
3	1	3	5	3	17
4	1	2	6	4	19
5	1	2	7	7	17
6	1	4	8	12	12
7	1	3	5	7	12
8	2	5	1	6	12
9	2	5	2	3	10
10	2	4	3	2	10
11	2	3	4	3	12
12	2	3	5	6	10
13	2	5	6	11	5
14	2	4	3	6	5
15	3	6	2	5	7
16	3	6	3	2	5
17	3	5	4	1	5
18	3	4	5	2	7
19	3	4	6	5	5
20	3	6	7	10	0
21	3	5	4	5	0
M_1		3.143	5.429	6.286	16.143
M_2		4.143	3.429	5.286	9.143
M_3		5.143	4.429	4.286	4.143
SD_1		0.899	1.718	3.039	2.968[b]
SD_2		0.899	1.718	3.039	2.968
SD_3		0.899	1.718	3.039	2.968

Covariance matrix for group 1 (n=7)

X1	.8095			
X2	-.5714	2.9524		
X3	.9524	2.8571	9.2381	
X4	-.6905	-2.4048	-5.8810	8.8095[b]

Covariance matrix for group 2 (n=7)

X1	.8095			
X2	-.5714	2.9524		
X3	.9524	2.8571	9.2381	
X4	-.6905	-2.4048	-5.8810	8.8095

Covariance matrix for group 3 (n=7)

X1	.8095			
X2	-.5714	2.9524		
X3	.9524	2.8571	9.2381	
X4	-.6905	-2.4048	-5.8810	8.8095

Pooled within-groups covariance matrix (n=21)[c]

X1	.8095			
X2	-.5714	2.9524		
X3	.9524	2.8571	9.2381	
X4	-.6905	-2.4048	-5.8810	8.8095

aera9801.wk1 3/20/98

[a]The "response variables" in a discriminant analysis are the intervally-scaled variables. In a *DDA* the response variables are the intervally-scaled criterion variables being predicted by group membership data. In a *PDA* the response variables are the intervally-scaled predictor variables predicting group membership.

[b]The variance on the diagonal of the variance/covariance matrix is the square of the SD of the variable (e.g., $2.968^2 = 8.8095$), and the SD of the variable is the square root of the variance of the variable (e.g., $8.8095^{.5} = 2.968$).

[c]Because here the group sizes are equal and the variance-covariance matrices computed seperately "within" each group are also exactly equal (staticians call this "homogeneity" of the covariance matrices--it sounds more sophisticated than simply [clearly] saying these matrices are equal), the weighted average of the covariance matrices (called the "pooled" covariance matrix) also equals each of the three separate group covariance matrices.

Table 13
Heuristic *Results* Illustrating
the *Context Specificity* of GLM Weights

Weights in the Context of 3 Response Variables

Standardized canonical discriminant function coefficients

	Func 1	Func 2
X1	1.50086	-.01817
X2	1.25012	1.16078
X3	-1.37261	-.44995

Structure matrix

	Func 1	Func 2
X1	.56076	-.60392*
X2	-.05557	.92134*
X3	-.16600	.17877*

Weights in the Context of 4 Response Variables

Standardized canonical discriminant function coefficients

	Func 1	Func 2
X1	-.47343	1.22249
X2	-.12685	1.77579
X3	1.09588	-1.04760
X4	1.16456	.56180

Structure matrix

	Func 1	Func 2
X1	-.34600*	.05602
X2	.09855	.48590*
X3	.10242*	-.01658
X4	.63238*	.09130

Table 14
Heuristic *Data Set #1* Illustrating
the *Importance of Both Function and Structure Coefficients*

ID/ Stat.	Grp	Response Variable* X1	X2	X3
1	1	0	13	13
2	1	4	6	18
3	1	2	9	33
4	1	6	4	8
5	1	8	3	13
6	1	10	3	25
7	1	12	4	30
8	1	14	6	20
9	1	18	13	25
10	1	16	9	5
11	2	1	14	9
12	2	5	7	14
13	2	3	10	29
14	2	7	5	4
15	2	9	4	9
16	2	11	4	21
17	2	13	5	26
18	2	15	7	16
19	2	19	14	21
20	2	17	10	1
21	3	3	11	10
22	3	7	4	15
23	3	5	7	30
24	3	9	2	5
25	3	11	1	10
26	3	13	1	22
27	3	15	2	27
28	3	17	4	17
29	3	21	11	22
30	3	19	7	2
M_1		9.000	7.000	19.000
M_2		10.000	8.000	15.000
M_3		12.000	5.000	16.000
SD_1		5.745	3.633	8.832
SD_2		5.745	3.633	8.832
SD_3		5.745	3.633	8.832

aera9803.wk1 3/21/98

*The "response variables" in a discriminant analysis are the intervally-scaled variables. In a *DDA* the response variables are the intervally-scaled criterion variables being predicted by group membership data. In a *PDA* the response variables are the intervally-scaled predictor variables predicting group membership.

Table 15
Heuristic *Results #1* Illustrating
the *Importance of Both Function and Structure Coefficients*

STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

	FUNC 1	FUNC 2
X1	-0.50132	-0.42337
X2	0.86161	-0.32427
X3	**0.07938**	0.84594

STRUCTURE MATRIX

	FUNC 1	FUNC 2
X1	-0.50132*	-0.42337
X2	0.86161*	-0.32427
X3	**0.07938**	0.84594*

Table 16
Heuristic *Data Set #2* Illustrating
the *Importance of Both Function and Structure Coefficients*

ID/ ID	Grp	Response Variable X1	X2	X3
1	1	29.504	42.923	29.576
2	1	35.377	40.427	37.666
3	1	38.646	30.333	29.319
4	1	32.166	29.527	29.132
5	1	42.123	37.132	37.234
6	1	53.744	28.508	35.073
7	1	32.359	49.590	44.558
8	1	36.474	44.465	29.162
9	1	48.948	38.320	41.963
10	1	50.738	39.708	41.576
11	1	50.535	39.256	49.973
12	1	54.179	41.307	45.286
13	1	33.117	40.453	33.975
14	1	31.286	43.078	31.644
15	1	49.303	45.602	54.567
16	1	39.003	53.124	48.315
17	1	40.929	45.935	37.544
18	1	47.503	42.345	39.500
19	1	60.888	25.192	41.614
20	1	49.430	47.577	48.575
21	1	34.541	45.680	33.675
22	1	57.003	44.198	50.022
23	1	49.220	30.174	41.746
24	1	26.350	61.440	47.090
25	1	23.518	59.255	39.290
26	1	34.368	46.340	32.676
27	1	54.078	39.067	49.668
28	1	60.099	39.284	47.905
29	1	44.404	50.167	49.101
30	1	48.057	53.530	53.324
31	1	38.759	49.947	35.416
32	1	47.351	42.744	33.542
33	1	44.271	51.256	41.813
34	1	54.653	56.111	57.113
35	1	50.281	29.187	40.420
36	1	42.553	53.080	46.333
37	1	50.407	51.156	46.945
38	1	43.467	44.039	39.270
39	1	35.070	58.895	54.395
40	1	50.100	42.617	44.246
41	2	47.031	39.315	42.967
42	2	34.440	39.074	28.846
43	2	58.123	28.294	48.125
44	2	38.938	51.302	44.884
45	2	51.361	50.766	51.048

46	2	62.011	37.532	53.917
47	2	51.508	38.750	45.354
48	2	44.008	59.607	52.539
49	2	42.859	54.788	50.448
50	2	45.016	49.405	39.081
51	2	42.240	58.822	55.714
52	2	38.303	63.250	50.925
53	2	46.990	55.456	59.208
54	2	59.590	49.553	54.338
55	2	59.283	44.703	57.759
56	2	61.041	36.092	49.032
57	2	50.987	47.080	50.649
58	2	40.836	47.054	40.398
59	2	48.792	53.000	54.324
60	2	69.031	36.061	54.631
61	2	46.701	50.368	48.506
62	2	47.542	38.375	34.157
63	2	45.837	67.118	55.424
64	2	56.153	47.024	56.497
65	2	37.357	69.471	56.009
66	2	55.308	48.748	51.137
67	2	52.835	54.825	56.223
68	2	53.123	47.373	49.864
69	2	47.460	57.273	51.987
70	2	54.667	55.468	54.064
71	2	51.908	63.035	63.285
72	2	46.339	54.534	48.788
73	2	53.074	49.063	51.585
74	2	43.738	49.927	45.095
75	2	55.794	54.218	59.348
76	2	55.442	60.159	59.143
77	2	51.622	57.054	55.261
78	2	48.822	47.487	49.094
79	2	62.728	45.031	56.524
80	2	68.232	44.920	56.643
81	3	40.712	56.076	48.200
82	3	63.342	41.788	52.130
83	3	45.558	62.146	60.511
84	3	36.825	50.628	38.987
85	3	51.950	37.143	44.118
86	3	66.036	55.171	55.063
87	3	54.325	54.081	57.507
88	3	50.350	63.482	61.656
89	3	60.064	57.666	62.915
90	3	46.860	53.651	56.202
91	3	49.541	60.881	48.760
92	3	48.559	67.626	59.086
93	3	53.106	59.407	60.727
94	3	45.581	47.690	42.188
95	3	55.511	65.603	61.156
96	3	50.678	50.874	48.134
97	3	56.552	57.215	60.662
98	3	51.219	52.479	54.457

99	3	65.509	45.815	61.583
100	3	47.918	64.920	56.355
101	3	61.419	52.850	62.042
102	3	44.501	58.930	53.558
103	3	74.096	47.502	57.292
104	3	46.979	59.091	47.819
105	3	51.521	51.319	47.794
106	3	49.607	61.195	57.117
107	3	61.220	59.648	64.766
108	3	61.227	51.405	51.763
109	3	50.665	67.000	73.107
110	3	55.987	62.945	65.147
111	3	66.710	40.179	51.533
112	3	60.045	54.641	64.535
113	3	63.249	47.013	57.722
114	3	56.559	69.593	64.217
115	3	73.365	47.814	53.417
116	3	57.605	63.250	66.086
117	3	51.241	69.792	52.887
118	3	65.732	60.535	67.970
119	3	58.946	69.314	64.397
120	3	74.402	49.980	74.818

Table 17
Heuristic *Results #2* Illustrating
the *Importance of Both Function and Structure Coefficients*

STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

	FUNC 1	FUNC 2
X1	0.93660	1.07729
X2	0.95259	1.43338
X3	**-0.05507**	-1.70996

STRUCTURE MATRIX

	FUNC 1	FUNC 2
X1	0.54141*	-0.28008
X2	0.56453*	0.24316
X3	**0.81431***	-0.55744

Table 18
Heuristic *Data Set #3* Illustrating
the *Importance of Both Function and Structure Coefficients*

ID/		Response Variable		
ID	Grp	X1	X2	X3
1	1	31.107	41.920	44.130
2	1	37.386	43.111	55.702
3	1	40.301	29.292	40.991
4	1	32.981	21.197	33.741
5	1	43.659	38.767	49.266
6	1	56.148	36.705	51.915
7	1	33.099	46.916	53.419
8	1	37.419	42.930	36.720
9	1	50.786	44.536	56.564
10	1	52.673	47.540	56.989
11	1	52.384	46.404	65.522
12	1	56.199	51.374	61.670
13	1	33.543	33.607	37.093
14	1	31.563	35.508	33.774
15	1	50.840	52.470	68.917
16	1	39.741	53.984	57.161
17	1	41.753	45.846	44.518
18	1	48.821	46.360	49.784
19	1	63.104	34.040	55.862
20	1	50.748	53.881	60.467
21	1	34.687	39.388	34.741
22	1	58.809	55.013	64.993
23	1	50.412	30.512	48.941
24	1	25.678	52.286	45.763
25	1	22.628	47.189	35.085
26	1	34.290	38.953	31.661
27	1	55.465	44.953	60.297
28	1	61.929	49.927	61.620
29	1	45.002	51.401	55.457
30	1	48.912	58.554	62.578
31	1	38.876	46.354	36.464
32	1	48.112	43.765	37.997
33	1	44.788	52.322	46.752
34	1	55.896	66.513	69.875
35	1	51.101	27.581	43.654
36	1	42.750	52.288	49.343
37	1	51.167	55.825	53.826
38	1	43.702	41.100	40.248
39	1	34.634	53.552	54.444
40	1	50.778	44.174	48.748
41	2	47.442	37.269	44.752
42	2	33.891	26.956	21.808
43	2	59.232	30.951	53.661
44	2	38.516	45.224	42.184
45	2	51.874	54.318	55.888

46	2	63.316	45.484	62.852
47	2	52.006	38.689	47.471
48	2	43.839	59.299	53.661
49	2	42.588	52.051	49.779
50	2	44.862	46.513	37.011
51	2	41.847	56.389	55.251
52	2	37.563	58.752	47.927
53	2	46.886	55.382	60.616
54	2	60.409	57.514	61.021
55	2	60.062	50.884	63.752
56	2	61.939	41.016	53.500
57	2	51.123	47.324	51.388
58	2	40.200	39.241	34.206
59	2	48.717	53.027	54.776
60	2	70.405	46.587	63.022
61	2	46.339	47.253	45.489
62	2	47.215	32.182	27.762
63	2	45.367	68.085	54.977
64	2	56.457	50.116	58.726
65	2	36.227	64.334	51.076
66	2	55.522	51.540	52.288
67	2	52.817	57.229	57.121
68	2	53.035	47.277	48.144
69	2	46.942	55.647	48.891
70	2	54.621	58.601	54.281
71	2	51.618	66.050	63.898
72	2	45.624	50.589	43.209
73	2	52.831	48.600	48.886
74	2	42.785	42.318	36.503
75	2	55.743	57.369	59.765
76	2	55.355	64.758	60.197
77	2	51.246	57.704	53.225
78	2	48.200	42.878	42.936
79	2	63.135	50.564	58.436
80	2	69.042	54.695	61.503
81	3	39.424	47.331	38.259
82	3	63.734	46.500	52.782
83	3	44.596	58.825	55.308
84	3	35.115	36.477	23.724
85	3	51.373	30.868	35.705
86	3	66.467	65.097	58.124
87	3	53.858	54.360	54.014
88	3	49.569	63.348	57.775
89	3	59.997	63.364	63.516
90	3	45.747	47.449	47.539
91	3	48.617	58.883	41.743
92	3	47.527	66.677	53.470
93	3	52.395	59.477	56.346
94	3	44.285	38.227	29.425
95	3	54.960	69.305	58.946
96	3	49.757	46.332	39.322
97	3	56.022	58.921	57.115
98	3	50.216	48.169	45.791

99	3	65.529	50.497	60.166
100	3	46.575	61.215	47.174
101	3	60.970	55.521	58.002
102	3	42.729	49.800	39.596
103	3	74.471	57.862	57.693
104	3	45.310	51.507	33.811
105	3	50.109	44.526	34.282
106	3	48.046	55.814	45.142
107	3	60.504	62.806	59.641
108	3	60.510	52.098	43.801
109	3	49.050	63.455	63.262
110	3	54.709	62.040	56.101
111	3	66.204	40.738	42.938
112	3	58.976	53.925	55.458
113	3	62.231	45.499	46.641
114	3	54.781	68.140	51.342
115	3	72.817	52.917	44.863
116	3	55.839	60.360	52.383
117	3	48.707	62.249	32.418
118	3	64.259	61.493	55.595
119	3	56.829	66.812	47.966
120	3	73.147	52.237	62.332

Table 19
Heuristic *Results #3* Illustrating
the *Importance of Both Function and Structure Coefficients*

STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

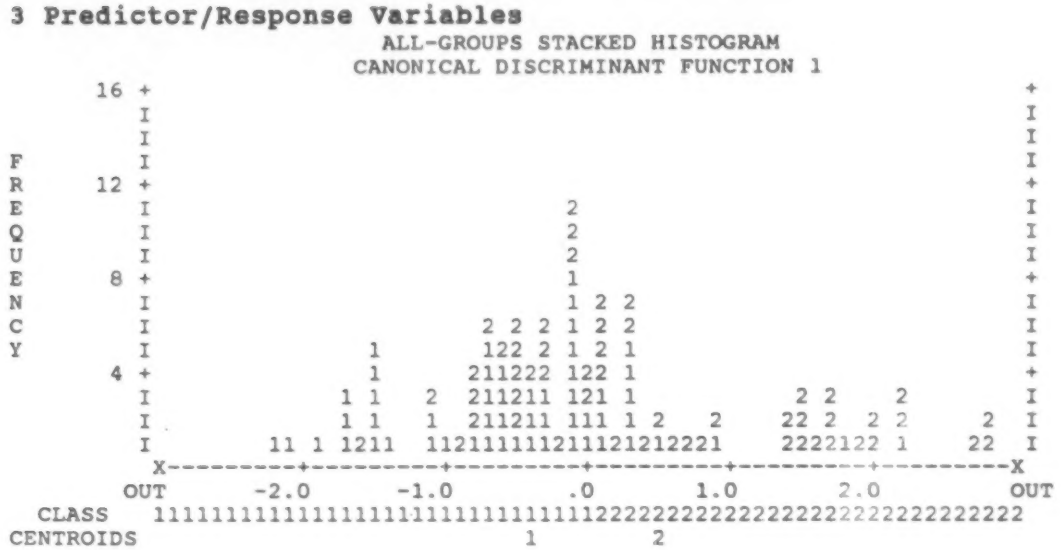
	FUNC 1	FUNC 2
X1	1.22956	0.28470
X2	1.21174	-0.20978
X3	**-1.58393**	0.89694

STRUCTURE MATRIX

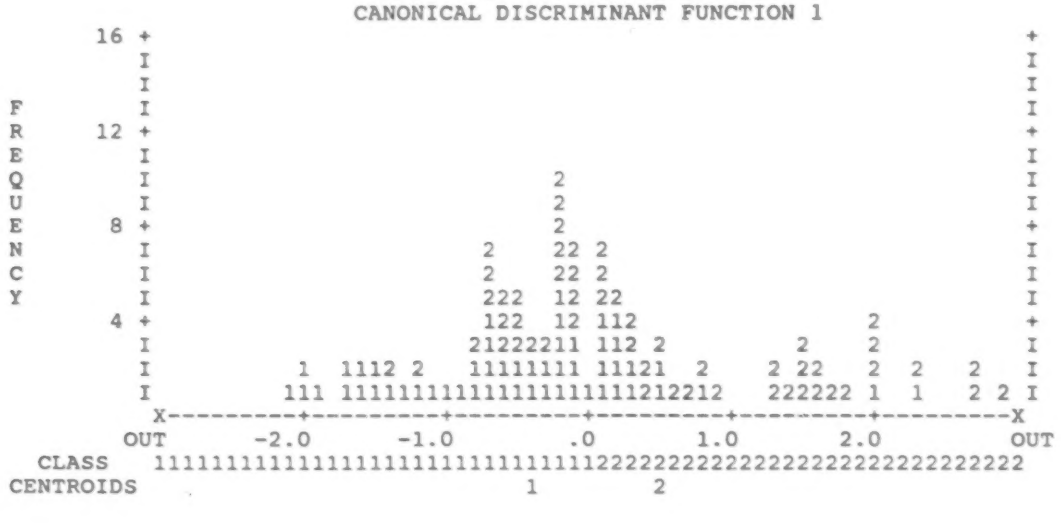
	FUNC 1	FUNC 2
X1	0.39129	0.82637*
X2	0.38294	0.39748*
X3	**-0.03464**	0.94557*

Figure 1
PDA Territorial Maps for the Table 5 Heuristic Data
Illustrating That *More* Predictors
May Actually *Hurt* Classification Accuracy

3 Predictor/Response Variables



4 Predictor/Response Variables



Note. Although the DDA effect size always stays the same or gets better (i.e., smaller) as more response variables are added (for these data, $\lambda_3 = 0.8094909$ while $\lambda_4 = 0.8050684$), the PDA hit rate can get worse as response variables are added.

APPENDIX A
SPSS/LISREL Program Illustraing That SEM is the Most General Case of the General Linear Model Using the Holzinger and Swineford (1939) Data

```
TITLE 'CANLISRL.SPS  Holzinger & Swineford (1939) Data **'.
COMMENT ************************************************************************.
COMMENT Holzinger, K.J., & Swineford, F. (1939). A study in factor analysis:.
COMMENT   The stability of a bi-factor solution (No. 48). Chicago, IL:.
COMMENT   University of Chicago. (data on pp. 81-91).
COMMENT ************************************************************************.
SET BLANKS=SYSMIS UNDEFINED=WARN.
DATA LIST
  FILE=abc FIXED RECORDS=2 TABLE
  /1 id 1-3 sex 4-4 ageyr 6-7
  agemo 8-9 t1 11-12 t2 14-15 t3 17-18 t4 20-21 t5 23-24 t6 26-27 t7 29-30 t8
  32-33 t9 35-36 t10 38-40 t11 42-44 t12 46-48 t13 50-52 t14 54-56 t15 58-60
  t16 62-64 t17 66-67 t18 69-70 t19 72-73 t20 74-76 t21 78-79 /2 t22 11-12
  t23 14-15 t24 17-18 t25 20-21 t26 23-24 .
EXECUTE.
COMPUTE SCHOOL=1.
IF (ID GT 200)SCHOOL=2.
IF (ID GE 1 AND ID LE 85)GRADE=7.
IF (ID GE 86 AND ID LE 168)GRADE=8.
IF (ID GE 201 AND ID LE 281)GRADE=7.
IF (ID GE 282 AND ID LE 351)GRADE=8.
IF (ID GE 1 AND ID LE 44)TRACK=2.
IF (ID GE 45 AND ID LE 85)TRACK=1.
IF (ID GE 86 AND ID LE 129)TRACK=2.
IF (ID GE 130)TRACK=1.
PRINT FORMATS SCHOOL TO TRACK(F1.0).
VALUE LABELS SCHOOL(1)PASTEUR (2) GRANT-WHITE/
  TRACK (1)JUNE PROMOTIONS (2)FEB PROMOTIONS/.
VARIABLE LABELS T1 VISUAL PERCEPTION TEST FROM SPEARMAN VPT, PART III
  T2 CUBES, SIMPLIFICATION OF BRIGHAM'S SPATIAL RELATIONS TEST
  T3 PAPER FORM BOARD--SHAPES THAT CAN BE COMBINED TO FORM A TARGET
  T4 LOZENGES FROM THORNDIKE--SHAPES FLIPPED OVER THEN IDENTIFY TARGET
  T5 GENERAL INFORMATION VERBAL TEST
  T6 PARAGRAPH COMPREHENSION TEST
  T7 SENTENCE COMPLETION TEST
  T8 WORD CLASSIFICATION--WHICH WORD NOT BELONG IN SET
  T9 WORD MEANING TEST
  T10 SPEEDED ADDITION TEST
  T11 SPEEDED CODE TEST--TRANSFORM SHAPES INTO ALPHA WITH CODE
  T12 SPEEDED COUNTING OF DOTS IN SHAPE
  T13 SPEEDED DISCRIM STRAIGHT AND CURVED CAPS
  T14 MEMORY OF TARGET WORDS
  T15 MEMORY OF TARGET NUMBERS
  T16 MEMORY OF TARGET SHAPES
  T17 MEMORY OF OBJECT-NUMBER ASSOCIATION TARGETS
  T18 MEMORY OF NUMBER-OBJECT ASSOCIATION TARGETS
  T19 MEMORY OF FIGURE-WORD ASSOCIATION TARGETS
  T20 DEDUCTIVE MATH ABILITY
  T21 MATH NUMBER PUZZLES
  T22 MATH WORD PROBLEM REASONING
  T23 COMPLETION OF A MATH NUMBER SERIES
  T24 WOODY-MCCALL MIXED MATH FUNDAMENTALS TEST
  T25 REVISION OF T3--PAPER FORM BOARD
  T26 FLAGS--POSSIBLE SUBSTITUTE FOR T4 LOZENGES.
SUBTITLE 'CCA ##############'.
correlations variables=t6 t7 t2 t4 t20 t21 t22/
  statistics=all .
manova t6 t7 with t2 t4 t20 t21 t22/
```

```
  print=signif(multiv eigen dimenr)/
  discrim=stan cor alpha(.999)/design .

SUBTITLE 'Function I   2nd Variate n=301 v=7'.
execute .
PRELIS
  /VARIABLES
  t2 (CO) t4 (CO) t20 (CO) t21 (CO) t22 (CO)
  t6 (CO) t7 (CO)
  /TYPE=CORRELATION
  /MATRIX=OUT(CR1)
LISREL
  /"1b  First Function   n=301 v=7"
  /DA NI=7 NO=301 MA=KM
  /MATRIX=IN(CR1)
  /MO BE=ZE PS=ZE TD=ZE LX=ID LY=FU,FI TE=SY,FR
   GA=FU,FI PH=SY,FR NX=2 NY=5 NK=2 NE=1
  /VA 1.0 PH(1,1) PH(2,2)
  /VA 1.0 LY(1,1)
  /FR LY(2,1) LY(3,1) LY(4,1) LY(5,1)
  /FR GA(1,1) GA(1,2)
  /OU SS FS SL=1 TM=1200 ND=5
SUBTITLE 'Function I   1st Variate n=301 v=7'.
execute .
PRELIS
  /VARIABLES
  t6 (CO) t7 (CO)
  t2 (CO) t4 (CO) t20 (CO) t21 (CO) t22 (CO)
  /TYPE=CORRELATION
  /MATRIX=OUT(CR2)
LISREL
  /"1a  First Function   n=301 v=7"
  /DA NI=7 NO=301 MA=KM
  /MATRIX=IN(CR2)
  /MO BE=ZE PS=ZE TD=ZE LX=ID LY=FU,FI TE=SY,FR
   GA=FU,FI PH=SY,FR NX=5 NY=2 NK=5 NE=1
  /VA 1.0 PH(1,1) PH(2,2) PH(3,3) PH(4,4) PH(5,5)
  /VA 1.0 LY(1,1)
  /FR LY(2,1)
  /FR GA(1,1) GA(1,2) GA(1,3) GA(1,4) GA(1,5)
  /OU SS FS SL=1 TM=1200 ND=5

SUBTITLE 'Function II  2nd Variate n=301 v=7'.
execute .
LISREL
  /"2b  Second Function  n=301 v=7"
  /DA NI=7 NO=301 MA=KM
  /MATRIX=IN(CR1)
  /MO BE=ZE PS=ZE TD=ZE LX=ID LY=FU,FI TE=SY,FR
   GA=FU,FI PH=SY,FR NX=2 NY=5 NK=2 NE=2
  /VA 1.0 PH(1,1) PH(2,2)
  /VA 1.0 LY(1,1) LY(1,2)
  /VA 0.76757 LY(2,1)
  /VA 2.34225 LY(3,1)
  /VA 2.13559 LY(4,1)
  /VA 3.17417 LY(5,1)
  /FR LY(2,2) LY(3,2) LY(4,2) LY(5,2)
  /VA 0.06992 GA(1,1)
  /VA 0.09682 GA(1,2)
  /FR GA(2,1) GA(2,2)
  /OU SS FS SL=1 TM=1200 ND=5
SUBTITLE 'Function II  1st Variate n=301 v=7'.
execute .
LISREL
  /"2a  Second Function  n=301 v=7"
  /DA NI=7 NO=301 MA=KM
```

```
/MATRIX=IN(CR2)
/MO BE=ZE PS=ZE TD=ZE LX=ID LY=FU,FI TE=SY,FR
 GA=FU,FI PH=SY,FR NX=5 NY=2 NK=5 NE=2
/VA 1.0 PH(1,1) PH(2,2) PH(3,3) PH(4,4) PH(5,5)
/VA 1.0 LY(1,1) LY(1,2)
/VA 1.05093 LY(2,1)
/FR LY(2,2)
/VA -.00729 GA(1,1)
/VA -.09934 GA(1,2)
/VA 0.16926 GA(1,3)
/VA 0.13288 GA(1,4)
/VA 0.36285 GA(1,5)
/FR GA(2,1) GA(2,2) GA(2,3) GA(2,4) GA(2,5)
/OU SS FS SL=1 TM=1200 ND=5
```

APPENDIX B
SPSS *Program* for the Table 5 Actual Data Illustrating That *More* Predictors May Actually *Hurt* Classification Accuracy

```
TITLE 'AERA9804.SPS  Holzinger & Swineford (1939) Data ****'
COMMENT*************************************************************************
COMMENT    Holzinger, K.J., & Swineford, F. (1939). A study in factor analysis:
COMMENT      The stability of a bi-factor solution (No. 48). Chicago, IL:
COMMENT      University of Chicago. (data on pp. 81-91)
COMMENT*************************************************************************
DATA LIST FILE=BT RECORDS=2
  /1 ID 1-3 SEX 4 AGEYR 6-7 AGEMO 8-9
  T1 11-12 T2 14-15 T3 17-18 T4 20-21 T5 23-24 T6 26-27
  T7 29-30 T8 32-33 T9 35-36 T10 38-40 T11 42-44 T12 46-48
  T13 50-52 T14 54-56 T15 58-60 T16 62-64 T17 66-67
  T18 69-70 T19 72-73 T20 74-76 T21 78-79
  /2 T22 11-12 T23 14-15 T24 17-18
  T25 20-21 T26 23-24
COMPUTE SCHOOL=1
IF (ID GT 200)SCHOOL=2
IF (ID GE 1 AND ID LE 85)GRADE=7
IF (ID GE 86 AND ID LE 168)GRADE=8
IF (ID GE 201 AND ID LE 281)GRADE=7
IF (ID GE 282 AND ID LE 351)GRADE=8
IF (ID GE 1 AND ID LE 44)TRACK=2
IF (ID GE 45 AND ID LE 85)TRACK=1
IF (ID GE 86 AND ID LE 129)TRACK=2
IF (ID GE 130)TRACK=1
PRINT FORMATS SCHOOL TO TRACK(F1.0)
VALUE LABELS SCHOOL(1)PASTEUR (2) GRANT-WHITE/
  TRACK (1)JUNE PROMOTIONS (2)FEB PROMOTIONS/
VARIABLE LABELS T1 VISUAL PERCEPTION TEST FROM SPEARMAN VPT, PART III
  T2 CUBES, SIMPLIFICATION OF BRIGHAM'S SPATIAL RELATIONS TEST
  T3 PAPER FORM BOARD--SHAPES THAT CAN BE COMBINED TO FORM A TARGET
  T4 LOZENGES FROM THORNDIKE--SHAPES FLIPPED OVER THEN IDENTIFY TARGET
  T5 GENERAL INFORMATION VERBAL TEST
  T6 PARAGRAPH COMPREHENSION TEST
  T7 SENTENCE COMPLETION TEST
  T8 WORD CLASSIFICATION--WHICH WORD NOT BELONG IN SET
  T9 WORD MEANING TEST
  T10 SPEEDED ADDITION TEST
  T11 SPEEDED CODE TEST--TRANSFORM SHAPES INTO ALPHA WITH CODE
  T12 SPEEDED COUNTING OF DOTS IN SHAPE
  T13 SPEEDED DISCRIM STRAIGHT AND CURVED CAPS
  T14 MEMORY OF TARGET WORDS
  T15 MEMORY OF TARGET NUMBERS
  T16 MEMORY OF TARGET SHAPES
  T17 MEMORY OF OBJECT-NUMBER ASSOCIATION TARGETS
  T18 MEMORY OF NUMBER-OBJECT ASSOCIATION TARGETS
  T19 MEMORY OF FIGURE-WORD ASSOCIATION TARGETS
  T20 DEDUCTIVE MATH ABILITY
  T21 MATH NUMBER PUZZLES
  T22 MATH WORD PROBLEM REASONING
  T23 COMPLETION OF A MATH NUMBER SERIES
  T24 WOODY-MCCALL MIXED MATH FUNDAMENTALS TEST
  T25 REVISION OF T3--PAPER FORM BOARD
  T26 FLAGS--POSSIBLE SUBSTITUTE FOR T4 LOZENGES
subtitle '0  PDA with 3 Predictor Variables **n=301'
discriminant groups=grade(7,8)/
  variables=t13 t17 t22/analysis=t13 t17 t22/
  method=direct/priors=equal/save=scores(discrim)/
  classify=pooled/
  statistics=mean stddev gcov tcov corr boxm coef table/
  plot=all
```

```
select if (discrim1 lt -1.5 or discrim1 gt 1.5
  or (discrim1 gt -.3 and discrim1 lt .3))
sort cases by grade id
list variables=id grade t13 t17 t22 t16/
  cases=999/format=numbered
subtitle '1  PDA with 3 Predictor Variables **n=107'
discriminant groups=grade(7,8)/
  variables=t13 t17 t22/analysis=t13 t17 t22/
  method=direct/priors=equal/save=class(LDFCL3)/
  classify=pooled/
  statistics=mean stddev gcov tcov corr boxm coef table/
  plot=all
subtitle '2  PDA with 4 Predictor Variables **n=107'
discriminant groups=grade(7,8)/
  variables=t13 t17 t22 t16/analysis=t13 t17 t22 t16/
  method=direct/priors=equal/save=class(LDFCL4)/
  classify=pooled/
  statistics=mean stddev gcov tcov corr boxm coef table/
  plot=all
subtitle '3  Compare the 4 Sets of Classification Results!!'
compute lcf31=(T13 * 0.1091137) + (T17 * -0.06245298)
  + (T22 * 0.1659288) + -12.84927
compute lcf32=(T13 * 0.1117800) + (T17 * 0.06471948)
  + (T22 * 0.2171317) + -15.91867
compute lcf41=(T13 * -0.008489698) + (T17 * -0.5090838)
  + (T22 * -0.09004268) + (T16 * 1.974350) + -97.20442
compute lcf42=(T13 * -0.007301202) + (T17 * -0.3875236)
  + (T22 * -0.04205625) + (T16 * 1.999159) +  -102.4071 .
compute LCFCL3=8
if (lcf31 gt lcf32)LCFCL3=7
compute LCFCL4=8
if (lcf41 gt lcf42)LCFCL4=7
print formats LCFCL3 LCFCL4 (F1)
variable labels
  lcf31 'Linear Class Function (LCF) score #1 3 preds'
  lcf32 'Linear Class Function (LCF) score #2 3 preds'
  lcf41 'Linear Class Function (LCF) score #1 4 preds'
  lcf42 'Linear Class Function (LCF) score #2 4 preds'
  LCFCL3 'LCF classification  3 preds'
  LCFCL4 'LCF classification  4 preds'
  LDFCL3 'LDF classification  3 preds'
  LDFCL4 'LDF classification  4 preds'
list variables=id grade LDFCL3 LDFCL4 LCFCL3 LCFCL4/
  cases=9999/format=numbered
crosstabs grade by LDFCL3
crosstabs grade by LDFCL4
crosstabs grade by LCFCL3
crosstabs grade by LCFCL4
subtitle '1 LDFCL3 and LCFCL3 <>'
temporary
select if (LDFCL3 ne LCFCL3)
list variables=id grade LDFCL3 LDFCL4 LCFCL3 LCFCL4/cases=99
subtitle '2 LDFCL4 and LCFCL4 <>'
temporary
select if (LDFCL4 ne LCFCL4)
list variables=id grade LDFCL3 LDFCL4 LCFCL3 LCFCL4/cases=99
subtitle '3 LDFCL4 and LDFCL3 <>'
temporary
select if (LDFCL4 ne LDFCL3)
list variables=id grade LDFCL3 LDFCL4 LCFCL3 LCFCL4/cases=99
subtitle '4 LCFCL4 and LCFCL3 <>'
temporary
select if (LCFCL4 ne LCFCL3)
list variables=id grade LDFCL3 LDFCL4 LCFCL3 LCFCL4/cases=99
```

APPENDIX C
SPSS *Program* for the Table 9 Heuristic Data Illustrating that Stepwise Methods Do *Not* *Identify the Best* Variable Set

```
title 'AERA9802.SPS ************************************'.
data list file=abc records=1 table/
  ID grp x1 to x4 (2F4,4F9.3)
list variables=all/cases=99 .
subtitle '1  Stepwise DDA ##############################'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x1 to x4/
  method=wilks/maxsteps=2/
  statistics=mean stddev gcov cov boxm/ .
subtitle '2a Enter X1,X2  @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x1 x2/
  method=direct/ .
subtitle '2b X1,X2  Show 1-way MANOVA is DDA !!!!!!!!!!!!!'.
manova x1 x2 by grp(1,3)/print=signif(multiv eigen dimenr)/
  discrim(stan corr alpha(.999))/design .
subtitle '3a Enter X1,X3  @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x1 x3/
  method=direct/ .
subtitle '3b X1,X3  Show 1-way MANOVA is DDA !!!!!!!!!!!!!'.
manova x1 x3 by grp(1,3)/print=signif(multiv eigen dimenr)/
  discrim(stan corr alpha(.999))/design .
subtitle '4a Enter X1,X4  @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x1 x4/
  method=direct/ .
subtitle '4b X1,X4  Show 1-way MANOVA is DDA !!!!!!!!!!!!!'.
manova x1 x4 by grp(1,3)/print=signif(multiv eigen dimenr)/
  discrim(stan corr alpha(.999))/design .
subtitle '5a Enter X2,X3  @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x2 x3/
  method=direct/ .
subtitle '5b X2,X3  Show 1-way MANOVA is DDA !!!!!!!!!!!!!'.
manova x2 x3 by grp(1,3)/print=signif(multiv eigen dimenr)/
  discrim(stan corr alpha(.999))/design .
subtitle '6a Enter X2,X4  @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x2 x4/
  method=direct/ .
subtitle '6b X2,X4  Show 1-way MANOVA is DDA !!!!!!!!!!!!!'.
manova x2 x4 by grp(1,3)/print=signif(multiv eigen dimenr)/
  discrim(stan corr alpha(.999))/design .
subtitle '7a Enter X3,X4  @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@'.
discriminant
  groups=grp(1,3)/variables=x1 to x4/analysis=x3 x4/
  method=direct/ .
subtitle '7b X3,X4  Show 1-way MANOVA is DDA !!!!!!!!!!!!!'.
manova x3 x4 by grp(1,3)/print=signif(multiv eigen dimenr)/
  discrim(stan corr alpha(.999))/design .
```

APPENDIX D
SPSS for Windows *Program*
for the Table 12 Heuristic Data Illustrating the *Context Specificity* of GLM Weights

```
set printback=listing blanks=sysmis undefined=warn .
COMMENT  'AERA9801.SPS' .
title 'Illustrate **Context Specificity** of GLM Weights' .
data list
  file='c:\123\temp.prn' fixed records=1 table
  /1 id 1-3 grp 8 x1 14-15 x2 21-22 x3 28-29 x4 35-36 .
list variables=all/cases=9999/ .
subtitle '1  Discrim  ***Smaller Variable Set***' .
discriminant groups=grp(1,3)/variables=x1 to x3/
  analysis=x1 to x3/
  method=direct/priors=equal/save scores(dscr)/
  plot=cases/classify=pooled/
  statistics=mean stddev gcov cov corr boxm coef table.
variable label
  dscr1 'Discriminant score Func I   3 predictors'
  dscr2 'Discriminant score Func II  3 predictors' .
execute .
subtitle '2  Discrim  ###Larger Variable Set###' .
discriminant groups=grp(1,3)/variables=x1 to x4/
  analysis=x1 to x4/
  method=direct/priors=equal/save scores(dscore)/
  plot=cases/classify=pooled/
  statistics=mean stddev gcov cov corr boxm coef table.
variable label
  dscore1 'Discriminant score Func I   4 predictors'
  dscore2 'Discriminant score Func II  4 predictors' .
execute .
```

APPENDIX E
SPSS *Program* for the Heuristic Data (Tables 14, 16, and 18) Illustrating the *Importance of Both Function and Structure Coefficients*

```
title 'AERA9803.SPS ****************************************'.
data list file=abc records=1 table/
  ID 2-3 grp 8 x1 14-15 X2 21-22 X3 28-29
list variables=all/cases=99 .
subtitle '1  Uncorrelated Response Variables ##############'.
discriminant
  groups=grp(1,3)/variables=x1 to x3/analysis=x1 to x3/
  method=direct/
  statistics=mean stddev gcov cov boxm/ .
```